

AIS 2017

**12th International Symposium
on Applied Informatics and Related Areas
organized in the frame of
Hungarian Science Festival 2017
by Óbuda University**

PROCEEDINGS

November 9, 2017
Székesfehérvár, Hungary

ISBN 978-963-449-032-6

Adaptive Case Management and Dynamic Business Process Modeling

A proposal for document-centric and formal ap- proach

Bálint Molnár*, Khawla Bouafia**

*, ** *Information Systems Department, Eötvös Loránd University, ELTE, Budapest, Hungary*
molnarba@inf.elte.hu, bouafia@inf.elte.hu

Abstract—The modern Information Systems (IS) serve a dynamically changing organization environment with actual business processes (BPs) and documents in meat-space and with their representation within cyberspace. The challenge of the modeling is to track the continuous changes that bear on documents and flow structure of BPs.

The traditional approaches for dynamic process modeling take into account the design phase and operational phase within life cycles of both documents and BPs. However, the most recent management science approaches as agile company, lean management etc. enforce that the enhancements should start at the analysis phase and then the required modifications embodying in the structure of processes and organization should cascade through the design and operational phase in consistent and integrated way. As the documents and processes are strongly coupled to each other, a framework is needed that can reconcile the heterogenous views and models into a unified one and this framework should be grounded in a mathematically sound theory. The hypergraph as a mathematical structure is very flexible thereby it offers the opportunity for unified and uniform handling of models and providing solutions for representing and controlling of dynamic processes in the major “seasons” of their life cycles.

Keywords: IS modeling, document-centric modeling, IS architecture, Zachman framework, hypergraphs, business process (BP) modeling, dynamic process modeling.

I. INTRODUCTION

Most recently, the rapidly changing business environment and thereby originated adaptivity requirement against Business IS led to the agile business approach as a management science philosophy. The agility at enterprise level results in continuously changing BPs. The agility is enforced by trends in the organizational environment caused by market or authorities in time dimension. For that reason, a modeling framework that should track the changes should consider the requirements for functions in time dimensions. In an IS environment, the changes can be captured in the form of data items, data collections, documents and structure of processes. The models should grasp the various views and perspectives of the functioning enterprise exploiting the service of IS [27]. At the organization and strategic

planning level, the overarching demand for modifications appears then the business and system analyst should understand the essence of amendments, and then the analysts should create an adequate representation. The representation at business analysis and process level should take into account the demands for changes at the data perspective. As the requirement for changes at data perspective can be perceived as modification in the structure of documents, collections of data and database schemas. The intimate interrelationship between documents and BPs can assist to deduce the requirements for changes dynamically. The documents function as inputs and outputs for processes so that the changes in document and data structures can be used to identify the requirements for changes in BPs.

Consequently, the demand for structural modification in documents and collections of data makes necessary a discovery mechanism that pinpoints the models and their scopes where the enhancements should be accomplished while the consistency, integrity and security objectives should be sustained. This approach is in concert with agile enterprise management, agile system development and system operations methodologies.

As motivating example, we may use an e-public administration example where the documents containing description for legal procedural rules may change by law or by regulatory authorities [20]. The legal procedural rules contain prescriptions on the structures and variables /data fields of generic documents. The variables are valued and bounded to specific data items during processing till the specific document achieves the finalized, ultimately the ground document status[17][21]. A certain set of procedural rules are mapped onto rules embedded into intentional document types, another set of procedural rules can be organized into structures of generic BPs. The codified changes of legal procedural rules at public administration level initiate changes for intentional document types and in a coherent way for structure and components of BP. The BPs of public administration are supported by workflows as a kind of operationalized BPs with interaction of human roles.

In Section 2, we provide a brief overview of the related literature. In Section 3, we describe the formal and mathematical background that assist in representing documents, data structure and processes, moreover we describe the proposed approach, and then in Section 4 we conclude our research and discuss the results and the planned research in future.

II. LITERATURE REVIEW

The concept of dynamic processes and its definition is widely differing. Generally, the definitions refer to changes within the external and internal environment, and the consequences can be traced through adding, deleting, replacing components representing activities [1]. The alteration at organizational level will be realized at operational level, neither the cascading effects on elements of the process nor on other processes, and nor on the interrelationships between processes are discussed. Reference [1] follow a similar track, the changes can be accomplished at operational level on the structure of the workflow and instance of a single process, however the modifications of content and business logic included in the activities are not analyzed. Jain et al. [10] examine the impact of internal changes on BPs and the capability of the processes to adapt themselves to the changing environment; the externally initiated alterations are not considered. Hermosillo et al. [18] postulate that processes should be capable for dynamic adaptation to different scenarios, although the method for adaptation cannot be exhibited in detail. Mejia et al. [15] outlines an approach for dynamic adaptation based on Even-Condition-Action methodology, and proposes a rule-based approach, nonetheless the focus of the paper on a context-aware adaptation at operational level based on rules.

The use of semi-structured, active XML documents and a disciplined design approach are discussed in [14] to construct IS from the viewpoint of active documents. Another paper [2] presents a design methodology for a systematic design process to organize and maintain large amounts of data in a Web site based IS through a hypermedia design methodology. For the design of large-scale IS based on web applications and web services, Reference [24] contains a method.

Beside the different analysis and design model for IS that underlie of BP Models and workflow, the concept of architectural approach is that plays a relevant role that can be used in model creation. The various architectures for IS have been used to assist in understanding the relationship between the different perspectives, aspects, components and single models [5][23][27]. Zachman architectures developed for IS in enterprise, TOGAF is developed by Open Group for software systems within companies. TOGAF method contains two main parts: The Architecture Development Method (ADM) and the Foundation Architecture with generic functions/services on which specific architectures and building blocks can be built[5].

Joeris [12] proposed a document based approach for modeling control and data flow for business activities and data interchange among them. Wewers et al. [25] present a system that supports a framework for inter-

organizational, document oriented workflow. The alignment and fitting between BPs and organization can be analyzed on the base of ontologies and semantic approaches[13].

The artifact-centric BP model uses three basic concepts: artifact classes, tasks, and business rules [6][26]. The tasks handle the artifacts, the business rules govern which tasks should be triggered and which artifacts will be manipulated [8] [9].

The document-centric approach in concert with BP representation is analyzed in several papers. The models for data, documents and processes can be represented in unified framework based on graph-theoretical approach, the hypergraph as a lingua franca for models seems to be suitable for this purpose [16][17][19].

A. *The theoretical background for unified representation*

The BPs exist in a complex environment of IS. The representations of processes and interrelated elements enforces an enterprise architecture (EA) based approach with multitude of models. To sustain such a complex environment through the whole life cycle of IS, a formalized but flexible modeling method is required. The document centric approach with process-oriented perception provides an opportunity that can be exploited by formalized description that are grounded in mathematics.

B. *Documents as drivers for modeling BPs*

The data collection that is tightly coupled to the dynamic BP can be grasped by the concept of documents[17]. We should define a document type hierarchy. The basic approach is that, the evolution of the originally unbounded fields, variables within document can represent the status of the actual documents. We may imagine that there is an overarching document that includes any document types and data items in a concrete enterprise. The generic types consist of this overarching enterprise document. The generic document type can be modelled as semi-structured XML, by document object model approach, or any other appropriate way[11], thereby a generic document type includes classes of documents that can be generated by exploiting either structure or rules that are codified into the generic type. The document types that are created out of a generic document type can be instantiated into free documents that contains free variables, or unbounded fields that are not yet valued during manipulation of documents. The free documents can be perceived as free-tuples of tableaux or tableaux queries [3]. A specific type of documents is the intentional document types that are part of a generic document type, they compose hierarchy of document types, the specific property of intentional document types that there are rules that are attached to parts of the internal structure. The rules provide tools to modify and adapt the structure of the documents in concert with the actual task during the flow of activities within a BP.

C. *Hypergraphs as sound ground for modeling*

The hypergraphs, especially the generalized hypergraphs provide a flexible structure to describe complex relationships that can be explored among models during analysis and design of IS. A hypergraph is a pair (V, E) of

a finite set $V = \{v_1, \dots, v_n\}$ and a set E of nonempty subsets of V . The elements of V are called vertices; the elements of E are called edges. The generalized hypergraph allows that hyperedges as nodes can be included in other hyperedge but the contained hyperedge should be different from the container hyperedge.

Definition 1. “The concept of the directed hypergraphs is an ordered pair of vertices and hyperarcs that are directed hyperedges”, i. e. each hyperarc is an ordered pair that contains a *tail* $\overrightarrow{e_i^+}$ and a *head* $\overleftarrow{e_i^-}$ of [7].

$$\overrightarrow{e_i} = \left(\overleftarrow{e_i^-} = (e_i^+, i); \overleftarrow{e_i^-} = (i, e_i^-) \right), e_i^+ \subseteq V, e_i^- \subseteq V \quad (1)$$

The enterprise, BP, document and data architecture can be represented by a unified modeling framework. This approach offers the chance to capture the essential and critical modeling issues of dynamic BPs [17][21].

D. Modeling Dynamic BPs

The deployments into production environment of changes in BPs results in high failure rates. The cause of this phenomenon is attributed to the inability to predict the outcome of the exercise without implementing the change in the physical environment.

Definition 2. “BP is a collection of activities that takes kinds of input (one or more) and creates an output that is of value to the customer. Or a set of activities undertaken in a specific objective. The responsibility for execution of the activities (all or part) by an actor represents a role” [28][29].

Users are assigned to roles to scope of delegated power, and to a set of tasks to be executed. The role within an organization is an abstraction, in fact a group of people may belong to a single role to avoid any bottleneck and enable distribution of tasks. This objective is achievable through information and communication technology (ICT) and automatization of Information Management. A BP model consists of a set of models that connected through complex relationships having integrity, consistency, security and logical constraints associated to them. This perception of BP model corresponds to the principles of EA and analysis methods of IS. The above-mentioned hypergraph representation provides the opportunity for representing each single model with its cross-references and constraints in a flexible way that can be handled by tools grounded in graph-theory. This modeling environment yields the chance that the results of analysis may flow through the various layers of models till the implementation and operation season smoothly.

Definition 3. “BP model is an abstraction of the way how the working systems and individuals have to meet a business need that is described in a standard for business procedures, and a representation of knowledge and expertise of the practice of the profession”.

A BP model is a kind of map that governs the course of the job from beginning to the end.

- It is the basis for process improvement which allows a better understanding of the process.
- It serves as a basis for decision support, affects the decision on setting priorities objectives, serves as the basis for resources.

- It allows to anticipate changes and developments.
- Even if it does not provide all the answers, it provides a vision of the strategy followed by the company.
- Verification: process models are analyzed to find errors in systems or procedures (e. g. potential deadlocks).
- Performance analysis: techniques like simulation can be used to understand the factors influencing response times, service levels.
- Configuration: models can be used to configure a system.

E. Existing models for representing BP

Designers use models to represent BPs in a graphical way. Models require to formalize the process by using formal methods and visual languages. We will describe and analyze each relevant model from the literature that can be used for representation of BPs in a hypergraph.

The classical *Petri net* was invented by Carl Adam Petri in the sixties [31]. Since then it has been used to model and analyze all kinds of processes with applications ranging from communication protocols, hardware, and embedded systems to flexible manufacturing systems, user interaction. In the last two decades, the classical Petri net has been extended with color, time and hierarchy [30][31]. These extensions facilitate the modeling of complex processes where data and time are important factors.

Finite State Machines (FSM) is a well-known modeling method in the formal specification systems. A FSM is a behavioral model that contains a finite number of states. The states are called : the initial -state, end-state and other for representing transitions.

Unified Modeling Language (UML) [32] offers specialized diagrams (including diagrams of activity diagram, sequence diagram, class diagram, state charts etc.) each having a specific function. The UML Activity Diagrams can be used to model BPs [33], to model the logic of the use cases or user scenarios, or to model a participant of the business with the related business activities and business logic [34]. The UML Activity Diagrams can model the internal logic of complex operations. The UML Activity Diagrams are the object oriented (OO) equivalent to the data flow and flowchart diagrams that are used structured development methods.

Business Process Execution Language (BPEL) [35] is a standardized language for specifying the behavior of a BP based on interactions between a process and its service partners. It defines how multiple service interactions Web Service Description Language (WSDL) document through a set of operations and messages that can be dealt with. A BPEL process uses a set of variables to represent the messages exchanged between partners. They also represent the state of the BP. WSDL document through a set of operations and messages that can be dealt with. A BPEL process uses a set of variables to represent the messages exchanged between partners. They also represent the state of the business process.

Business Process Modeling Notation (BPMN) [36] used to model BPs and a basic ontology to represent domain information model entities. BPMN create a standard to bridge the gap between design and implementation of BPs.

TABLE I.

COMPARATIVE ANALYSIS OF BP MODELING APPROACH

Model	Criteria		
	Semantic	Syntactic	Structural
Petri-net	<p><i>Formal semantics:</i> in classical Petri net and several enhancements (color, time, hierarchy). <i>Operational semantics:</i> described in terms of tokens in places in of Petri net [38]. Petri nets have an exact mathematical definition of their execution semantics, with a well-developed mathematical theory for process analysis [29].</p>	<p>Petri net is triplet (P, T, F): -P is a finite set of places. -T a finite set of transitions (P ∩ T = ∅). -F ⊆ (P × T) ∪ (T × P) is a set of arcs (flow relation).</p>	<p>Graphical nature: supports the communication with end-users. [39]</p>
FSM	<p>A FSM is a restricted Turing machine where the head can only perform "read" operations, and always moves from left to right. [39]</p>	<p>5-tuple A= (M,Q,q0,F,R): • M: an input alphabet. • Q: finite set of states of A. • q0 in Q is the initial state. • F subset of Q is final states • R: Q×X is the transition function.</p>	<p>Graphical nature of FSM supports communication because it is easy to understand by users and describe reactive behavior and dynamic systems.</p>
BPMN	<p>Semantic analysis of BPMN models is hindered by the heterogeneity of its constructs and the lack of an unambiguous definition of the notation. Formal semantics of BPMN is that of Wong and Gibbons, which uses Communicating Sequential Processes (CSP) as the target formal model.</p>	<p>syntactic rules are comprehensively documented in tables throughout the BPMN standard specification, the actual semantics is only described in narrative form</p>	<p>Initially positioned as a modeling formalism and only informally defined, it has matured into a fully fledged BP modeling and execution language based on a comprehensive metamodel together with an associated graphical modeling notation and an execution semantics defining how BPMN processes should be enacted. [40]</p>
UML	<p>The UML semantics is described in an informal manner [40]</p>	<p>UML have an abstract syntax textual notations</p>	<p><i>UML activity</i> models is expressed in process modeling visual language, and allows to model enterprise from different perspectives. The <i>UML</i> class diagrams characterize the abstract syntax of the language</p>
BPEL	<p>A BPEL process is described in terms of XM (eXtensible Markup Language) [35]</p>	<p>BPEL uses an XML-based syntax based on XML depends to the activities of BPEL [35]</p>	<p>BPEL's activities executed in order (workflow): Basic activities: invoke, receive, reply, assign, throw, wait, empty. Structured activities Sequence, Switch, While, Pick, Flow, Repeat-Until For-Each, If-Else.</p>
Process Algebra	<p><i>Operational semantics:</i> describes systems evolution in terms of labelled transitions. It is relatively close to an abstract machine based view of computation and might be considered as a mathematical formalization of some implementation strategy. <i>A denotational semantics:</i> maps a language to some abstract model[28]. <i>Algebraic semantics:</i> algebraic laws are the basic axioms of an equational system, and process [41]</p>	<p>The basic component of process algebra is its syntax as determined by the well-formed combination of operators and more elementary terms. The syntax of a process algebra is the set of rules that define the combinations of symbols that are considered to be correctly structured programs in that language. [41]</p>	<p>The different operators used in process algebras will be described by relying on the so called structural operational semantic (SOS) approach[41]. The various approaches are CCS CSP and ACP</p>

BPMN provides a visual language in the form of graphical notation for defining BPs in a diagram. BPMN is a *defacto* standard language for describing BP, especially at the level of domain analysis and high-level system design. A growing number of process design, EA, and workflow automation tools provide modeling environments for BPMN.

In computer science, the *process algebras* (or *process calculi*) are a diverse family of related approaches for formally modeling concurrent systems. Process algebras provide a tool for the high-level description of interactions, communications, and synchronizations between a collection of independent agents or processes[37].

F. Models Comparison

There are numerous BP model in the literature. To see the similarities and the discrepancies we may use three viewpoints in comparison.

1. **Semantic:** taking into account the meaning of model and element labels and comments.

2. **Syntactic:** taking into account the types of modeling languages and language elements.

3. **Structural:** taking into account the model structure. From this table:

- Petri net has a simple structure and it is easy to analyze, simple Petri nets are good for testing the model. But low-level net is not suitable for performance analysis. To enable this functionality, we need to use time and color.
- Petri net is a “graphical and mathematical tool of modeling discrete event dynamic systems”[43]. which applied in simulation of discrete-event dynamic systems.
- FSM are easy to understand by users and popular and widespread tools it is suitable for describing reactive behavior and dynamic systems and also used for performance control services.
- UML Activity diagrams and BPMN are quite similar technologies and they are suitable for static modeling of BPs. But BPMN is more suitable and has much more representational power, this is because UML covers all layers of EA, while BPMN is specifically BP and

simulation and for executing the process models themselves by automating wherever it is possible the process steps. But it gives a static image, or drawing, of BP without simulation capability.

- UML is a visual language for OO modeling approaches. It is particularly used in software modeling.
- UML helps model classes connections and shows sequence flows, conditions of BPs in the enterprises.
- The strength of the BPEL is in the structure of its process based on an XML for standardize the integration of enterprise applications.

There are two kinds of models: *Dynamic* and *Static*, most of currently used enterprise modeling technologies can be considered as static. In real life scenarios BPs are not static. That's why demand for non-static models appears. The reference [45][44] contains comparison between static modeling and dynamic modeling:

Dynamic model facilitates the display of activities and flow of events within a process. The advantage of using dynamic modeling is that it enables the outcome of a changed process to be evaluated prior to it being implemented into the physical environment.

Static models have deterministic nature and are independent of process sequence, it may depend on the data collection and documents that are processed during the flow of information.

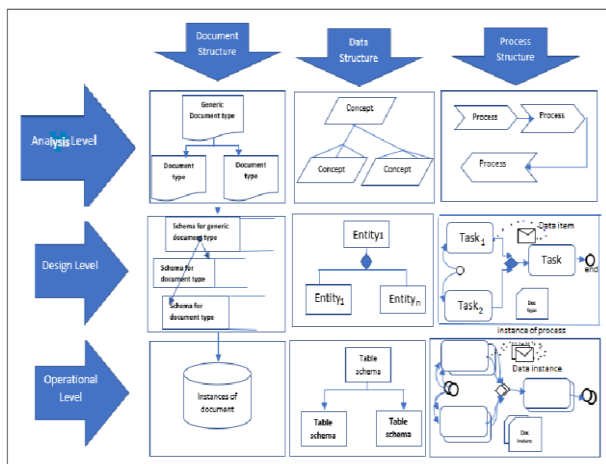


Figure 1 The architectural building blocks for a framework of unified modeling.

The main advantage of using static modeling technique is that it enables an in-depth understanding of the process being modelled. This approach includes the modeling of organizational structures, and of information carriers like documents or the modeling of relationships between business artifacts.

III. PROPOSED APPROACH FOR MODELING DYNAMIC PROCESSES

Enterprises either commercial businesses or government organizations are faced with a range of challenges recently. These challenges impact the architecture of these enterprises, also contentious changes in the environment such as changes in economy, society, physical environment, economics, culture and politics. For that reason, enterprises should

be able to focus their attention on all those impacts and finding ways to react a flexible way for the external stimuli that appear in the form of business events. The proposal is to extend the BP modeling approaches with an organization and planning level according to Zachman Framework [27] thereby an “*analysis season*” is created beside the “*design season*” and “*operational season*” [4].

In this approach, the data and business structure (operational level and design level) is expanded with an analysis phase (organizational and strategic planning level). The proposed framework is exhibited in Figure 1.

The analysis level provides the chances to observe and to detect business events that enforce changes in process flow, documents and document flow. The impact of business events influence both the structure of the workflow and the structure of document types.

The representation of BPs either in Petri-net or BPMN can be perceived as a document in XML format [46][45] [47]. The documents that are the media for data collections, and the process models can be represented using the concept of document types. The document types can be described meticulously in a hypergraph. The hyperedges can depict both the internal structure of document and BP models. Predicates of Description Logic and rules can be attached. to the intentional documents and to the relationships between model BPs, document types and documents [22]. The hypergraph and related mathematical tools provide the chance to keep and to enforce the consistency, integrity and compliance of models.

IV. CONCLUSION

The document types hierarchy and some basic Petri-nets described in XML are already represented in hypergraph. The basic consistency checking already is operational. The proposed formal modeling approach seems promising as the primary results of investigation and assessment that follows Design Research Science method demonstrates. There are efficiency issues that may be solved by other graph databases. The rapidly evolving software environment enforces to upgrade the underlying graph-database continuously and if the database becomes obsolete then exchange it to up-to-date graph-database that may provide representational capability for hypergraphs.

ACKNOWLEDGMENT

This work was supported by European Commission [grant number EFOP-3. 6. 3-VEKOP-16].

REFERENCES

- [1] Adams M. (2010). *Dynamic Workflow*, Hofstede et al. *Modern BP Automation*, Springer-Verlag Berlin Heidelberg, 123-145.
- [2] Atzeni, P. , Merialdo, P. , Mecca, G. , (2001). *Data-Intensive Web Sites: Design and Maintenance*, *World Wide Web*, 4, pp. 21–47.
- [3] Aho, A. V. , Sagiv, Y. , Ullman. , J. D. , (1979). Efficient optimization of a class of relational expressions. *ACM Trans. Database Syst.* 4, 4 (December 1979), 435-454. DOI=http://dx. doi. org/10. 1145/320107. 320112
- [4] Bell, M. (2008). *Service-oriented modeling (SOA): Service analysis, design, and architecture*. John Wiley & Sons.

- [5] Bent, H. V. D. , Sante, T. V. , Kerssens, D. , & Kemmeren, J. (2008). TOGAF, the open group architecture framework. Van Haren Publishing: Zaltbommel. <http://www.opengroup.org/togaf/>
- [6] Bhattacharya, K. , Gerede, C. , Hull, R. , Liu, R. and Su, J. (2007). Towards Formal Analysis of Artifact-Centric BP Models, in *BPM 2007*, LNCS, vol. 4714, G. Alonso, P. Dadam, M. Rosemann, Eds. , Heidelberg, Germany: Springer, pp. 288–304.
- [7] Bretto, A. (2013). *Hypergraph Theory: An Introduction*. Springer: Berlin, Heidelberg.
- [8] Hull, R. (2013). Data-Centricity and Services Interoperation in *International Conference on Service-Oriented Computing*, B. Samik, P. Cesare, Z. Liang, F. Xiang, Eds. , Berlin, Heidelberg, Germany: Springer, pp. 1-8
- [9] Hull, R. , (2008). Artifact-Centric BP Models: Brief Survey of Research Results and Challenges, in *On the Move to Meaningful Internet Systems: OTM 2008*, R. Meersman, Z. Tari, Eds. , Berlin, Heidelberg, Germany: Springer, pp 1152-1163.
- [10] Jain P. , Yeh P. Z. , Verma K. , Kass A. , Sheth A. (2008). Enhancing process-adaptation capabilities with web-based corporate radar technologies. In: *Proceedings of the first international workshop on Ontology-supported business intelligence*. ACM, 2-7.
- [11] Marini. , J. ,(2002). *Document Object Model (1 ed.)*. McGrawHill, Inc. , New York, NY.
- [12] Joeris, G. , (1997). Cooperative and integrated workflow and document management for engineering applications. In *Database and Expert Systems Applications*, Toulouse, France, pp. 68-73.
- [13] Kő, A. , Ternai, K. , (2011). A Development Method for Ontology Based BPs, in *eChallenges e-2011 Conference*, Florence, Italy.
- [14] Köppen, E. , & Neumann, G. (1999). Active hypertext for distributed web applications. In *Enabling Technologies: Infrastructure for Collaborative Enterprises*, 1999. (WET ICE'99) Proceedings. IEEE 8th International Workshops on (pp. 297-302). IEEE. DOI: 10. 1109/ENABL. 1999. 805216
- [15] Mejia Bernal J. F. , Falcarin P. , Morisio M. , Dai J. (2010). Dynamic context-aware BPs: a rule-based approach supported by pattern identification. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, 470-474.
- [16] Molnár, B. (2014). Applications of hypergraphs in informatics: a survey and opportunities for research. *Ann. Univ. Sci. Budapest. Sect. Comput*, 42, 261-282.
- [17] Molnár, B. , & Benczúr, A. (2013). Facet of modeling web IS from a document-centric view. *International Journal of Web Portals (IJWP)*, 5(4), 57-70. DOI: 10. 4018/ijwp. 2013100105
- [18] Hermosillo G. , Seinturier L. , Duchien L. (2010). Using complex event processing for dynamic business process adaptation. In: *Services Computing (SCC)*, 2010 IEEE International Conference on. IEEE, 466-473.
- [19] Molnár, B. , & Tarcsi, A. (2011, September). Architecture and system design issues of contemporary web-based IS In *Software, Knowledge Information, Industrial Management and Applications (SKIMA)*, 2011 5th International Conference on (pp. 1-8). IEEE. DOI: 10. 1109/SKIMA. 2011. 6089978
- [20] Molnár, B. , Béleczi, A. , & Benczúr, A. (2016a). Application of Legal Ontologies Based Approaches for Procedural Side of Public Administration. In *International Conference on Electronic Government and the IS Perspective* (pp. 135-149). Springer International Publishing.
- [21] Molnár, B. , Benczúr, A. , & Béleczi, A. (2016b). Formal approach to modelling of modern IS Int. J. Inf. Syst. Proj. Manag. (2016, to be published). <http://www.sciencesphere.org/ijispm/archive/ijispm-040404.pdf>
- [22] Molnár, B. , Benczúr, A. , & Béleczi, A. (2016, March). A Model for Analysis and Design of IS Based on a Document Centric Approach. In *Asian Conference on Intelligent Information and Database Systems* (pp. 290-299). Springer Berlin Heidelberg.
- [23] Noran, O. , (2005). A systematic evaluation of the C4ISR AF Using ISO 15704 Annex A (GERAM). *Computers in Industry*, 56, 407--427.
- [24] Rossi, G. , Schwabe, D. , Lyardet, F. (1999). Web application models are more than conceptual models, in: P. Chen et al. (Ed.), *Advances in Conceptual Modeling*, LNCS, vol. 1727, pp. 239-252, Springer-Verlag, Berlin
- [25] Wewers, T. , Wargitsch, C. (1998) Four dimensions of interorganizational, document-oriented workflow: a case study of the approval of hazardous-waste disposal. In: *System Sciences, The 31st Hawaii International Conference*, Hawaii, USA, vol. 4, pp. 332–341.
- [26] Yongchareon, S. and Liu, C. (2010). A Process View Framework for Artifact-Centric BPs, In *OTM 2010*. LNCS, vol. 6426. R. Meersman, T. S. Dillon, P. Herrero, Eds. , Heidelberg, Germany: Springer, pp. 26–43.
- [27] Zachman, J. A. (1987). A framework for IS architecture. *IBM systems journal*, 26(3), 276-292. DOI: 10. 1147/sj. 263. 0276.
- [28] Brahimi, M. , Bouzidi, L. (2008). Publiée une fois par année, la *Revue électronique suisse de science de l'information (RESSI)* . ResSI.
- [29] Zur Muehlen, M. (2004). *Workflow-based process controlling: foundation, design, and application of workflow-driven process IS (Vol. 6)*. Michael zur Muehlen.
- [30] P. Azema and G. Balbo. Verification of workflow nets. editors, *1248 of Lecture Notes in Computer Science:407–426*, 2006.
- [31] C. A. Petri. *Institut fur instrumentelle Mathematik*, Bonn, 1962. PhD thesis .
- [32] OMG. *UML : Superstructure UML Specification*, 2, 2005.
- [33] A. H. ter Hofstede ter Hofstede P. Wohed N. Russell, W. M. van der Aalst. On the suitability of uml 2. 0 activity diagrams for BP modelling. 53, 2006.
- [34] Wohed, P. , et al. (2005, October). Pattern-based analysis of the control-flow perspective of UML activity diagrams. In *International Conference on Conceptual Modeling* (pp. 63-78). Springer Berlin Heidelberg.
- [35] Jordan, D. , et al ,(2007). Web services BP execution language version 2. 0. OASIS standard, 11(120), 5.
- [36] Arkin, A. (2002). BP modeling language. BPML.org.
- [37] Baeten, J. C. (2005). A brief history of process algebra. *Theoretical Computer Science*, 335(2-3), 131-146.
- [38] Rozenberg, G. , & Reisig, W. (1998). *Lectures on Petri Nets I: Basic Models*. Lecture notes in computer science, 1491.
- [39] Fernández, et al. (2010). SBPMN—An easier BP modeling notation for business users. *Computer Standards & Interfaces*, 32(1), 18-28.
- [40] Arthur H. M. ter Hofstede Wil M. P. van der Aalst. *WORKFLOW PATTERNS The Definitive Guide*. Nick Russell, 2016.
- [41] De Nicola, R. (2014). A gentle introduction to process algebras. *Notes*, 7. .
- [42] G. D. Plotkin. A structural approach to operational semantics. *J. Log. Algebra. Program*, 17–139:60–61.
- [43] Salum, L. (2008). Petri nets and time modelling. *The International Journal of Advanced Manufacturing Technology*, 38(3), 377-382.
- [44] Rosenberg. A. , (2010) Dynamic versus static modeling types, *SAP Modeling Handbook - Modeling Standards*, <https://wiki.scn.sap.com/wiki/display/ModHandbook/SAP+Modeling+Handbook++Modeling+Standards> (accessed: April 27, 2017).
- [45] Billington, J. , et al. & Weber, M. (2003, June). The Petri net markup language: concepts, technology, and tools. In *International Conference on Application and Theory of Petri Nets* (pp. 483-505). Springer Berlin Heidelberg.
- [46] Weber, M. , & Kindler, E. (2003). The petri net markup language. In *Petri Net Technology for Communication-Based Systems* (pp. 124144). Springer Berlin Heidelberg.
- [47] Kalenkova, A. , De Leoni, M. , & van der Aalst, W. M. (2014). Discovering, Analyzing and Enhancing BPMN Models Using ProM. In *BPM (Demos)* (p. 36).

Fault-tolerant Software Solutions in Microcontroller Based Systems

György Györök, Bertalan Beszédes

* Óbuda University/ Alba Regia Technical Faculty, Székesfehérvár, Hungary
e-mail: {gyorok.gyorgy, bertalan.beszedes}@amk.uni-obuda.hu

Abstract—In this article, the focus is on the different kind of software solutions, how can a microcontroller increase the fault-tolerance level of the embedded system. At the beginning, it will be shown, the theoretical base connection between fault, error and failure, the possible causes of faults, fault tolerant solutions and fault tolerant software solutions. In the realization part, it will be shown, how the error-free running time and error detecting features can be increase the robustness of the system.

I. INTRODUCTION

Fault-masking architectures can be classified into mainly hardware or software categories. Of course, neither can stand alone. It is containing mainly the type of the solution, which is in its name. The duplication of frequently failing units (typically, power supply unit [1]) is the most common way to realize a hardware redundancy [2]. In software solutions, there are the multiple execution, the multiple measurements [3] and the majority voters as the most common major categories.

From the foregoing, it seems, depending to what kind of redundancy had been used, it has significantly impact to system performance, required power, weight, price and reliability. It is important to review the various methods to assess – the perspective of – the possibilities how to increase the reliability.

But first, let us declare the exact meanings the three basic concepts, fault, error and failure. These concepts are connected through causes and effects. The fault causes an error, which is causes failure.

A. Fault, error and failure

The fault is the errors proven or suspected cause. In case of a hardware component it could be short circuit, connection cut or parameter changes listed here, while in software case, it could be unexpected input combination, staying in an endless loop, make an addressing mistake, to mentioning only a few. An example, during manufacturing an AND gate, and the surface of the semiconductor had been polluted by a micro-sized dust particle, the AND gate's input may stay in high logic level.

The error – caused by the fault – is already appears the internal state of the device. [4] For example, if the AND gate's input gets a high logic level input voltage, the input signal is the same as the stacked leg's signal. If the input signal changing – gets a low logic level – the change is no longer transmitted through the input drive, and the AND gate's output will not change. That will cause an error in the system.

A failure occurs, when the error gets out of the system's output. The gate's output did not enforce in the logical function, so it affected for the system's output signals. Thereby, the error gets out to the outside world and become a failure. [5].

The three mentioned type of errors, appears in three different level (Fig. 1.). The fault is inherently physical, the error is modifying the internal state of the system, it's informational nature, and failure essentially affect to the outside world.

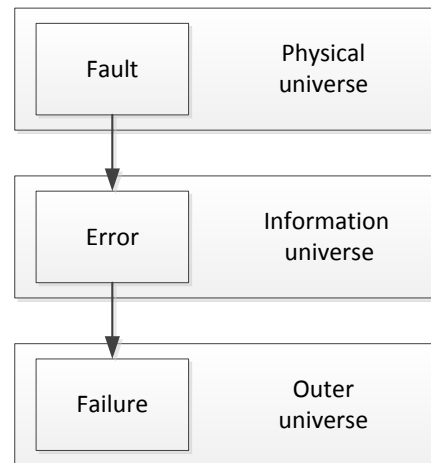


Figure 1. Three-universe

B. Causes of faults

The developing process of a device is starting with the specification phase. If this is not successful, that will cause a failure conception.

The formation of errors can be traced back to several things, like external interference or the consequences of the mistakes made during the design of the system or a component. The fault will not come to the surface, even during the examination of the final conformity test of the finished product. The error only turns out during the installation, operation and appears as a failure.

The topic of specification mistakes includes for example not correct timing, conversion, leveling, etc. between hardware or software modules.

The implementation mistakes may occurs when specification is implemented into practice. Improper or wrong component selection, not correct planning decisions or mistakes in coding may be the root cause.

The component imperfection is the most common source of fault. None of the components – neither from the same type – are matching perfectly, because their

parameters will may vary slightly. This can be easily remedied with conscious design or with component selection, however, the problems related to random component failures are much more significant. Typical causes are, for example, in case of microelectronic components, – within the case – the rupture of the bonds; metal corrosion; in case of electronic PCB's, some manufacturing imperfections or changes in the operating conditions – operating in extreme conditions. The failure type of component imperfection is also including the failure due to aging of the parts.

Strong emphasis should be placed on the control of external interferences. The foreseeing planning can give the ability to control these unpredictable effects. It can be classified into external interferences for example, the electromagnetic interference, radiation, the mistake of the operator, the result of a physical injury or environmental extremity (vibration, temperature, dust, humidity, ...), etc.

C. Nature of faults

If it could understand better the root causes of faults, then it could be developed improved procedures to prevent their formation. But until this, it need to be intervened after the appearance of an error, to maintain the operation of the system. [6]

As a first step, it is needed to know the types of faults:

- Source of faults:
 - Specification mistakes
 - Implementation mistakes
 - External interference
 - Component imperfection
- Type of faults:
 - Software
 - Hardware
 - Analog
 - Digital
- Duration of fault:
 - Permanent
 - Sporadic
 - Transient
- Expense of fault:
 - Local
 - Global
- Value of the fault:
 - Determined
 - Not determined

Very many fault combinations can be imagined based on the upper – broadly – classification. It is not expected to give a proper global solution for the problems. As a result, many theories have been advanced for the treatment of certain causes of errors.

The discussion of these is beyond the scope of the article, but it will be appreciated that, it needs to correct the relevant faults. After the exploration of the fault possibilities, it need to analyze the probability of the fault, and its consequences, and the resources spent for the troubleshooting.

It is practical – if the conditions allowing it – to choose the minimal hardware-intensive solutions, and prefer the software solutions, especially in embedded systems.

D. Propagation of faults

Fault tolerance, fault avoidance, fault prevention and fault masking is also a constructive technic, which can increase the reliability of the devices. Fig. 2. shows the application areas of the mentioned techniques.

Fault avoidance, fault prevention is the first defense line to prevent the formation of faults. Several techniques can be used, to increase the quality of the used components and the applied technologies. The method is characterized in that the disorder can be treated even before the formation. For example: plan criticism by an outside expert, using design practices to increasing reliability, component selection, oversizing, testing, shielding, and other methods to improve quality.

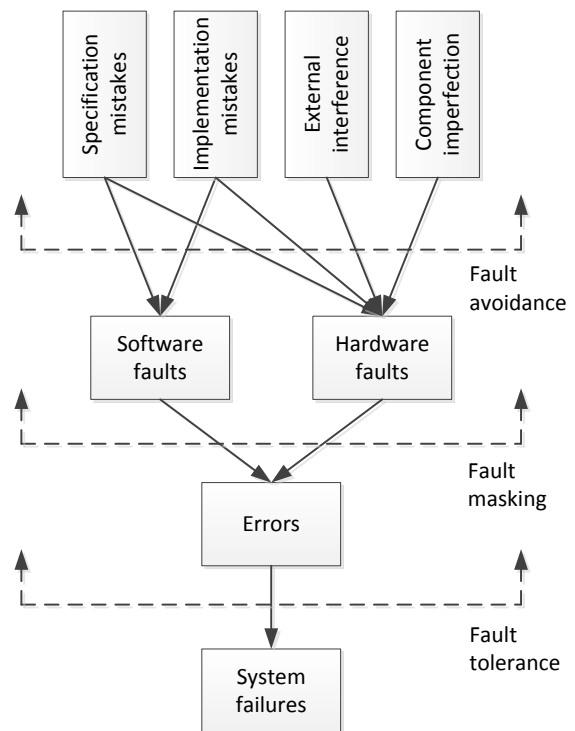


Figure 2. Preventing the spread of errors

The fault masking techniques are about to protect the system from the evolution of faults become an error. [7] When the fault has already occurred, fault masking is try to eliminate the fault's effect from the system – it does not let the fault to step out from the physical universe. A typical solution for a fault-masking system, is based on majority voting, where several autonomous decisions are made, and the result is given by the majority of voters. In this example, if one of the participants generates incorrect outcome, it becomes filter out, for this, it just need to be compared with the results of the other participants. Thus, the fault does not cause an error in the results.

The purpose of fault tolerance is to avoid faults, if the fault evolved an error. In this case fault masking and/or reconfiguration can be used. It can detect the error, then find out the source of it and the defective item will be

removed from the system and might set into operation a new one. [8]

II. FAULT-TOLERANT SOLUTIONS

At the beginning of the fault tolerance history, the usage of redundancy is always limited to physical hardware solutions. The most common solution to realize fault tolerance was to multiply the physical parts of the device, but nowadays we have more sophisticated solutions [9]. A redundant system – compared to simple system – have added information, resources or time.

The following types of redundancies are available:

- Hardware redundancy is when extra hardware is added to the system, it is typically used for fault detection and for fault tolerance. For example, multiplication of modules.
- Software redundancy means that, the added extra software modules, giving the possibilities to detect the faults – if possible, fix them –, next to the default functions of the original software. For example, timeout monitoring assigned to waiting's.
- Information redundancy is the extra information what is used to fault detection - if possible, fault correction – which would not be necessary for the default functions of the device. For example, using error correcting bits.
- Time redundancy is the extra time what is used to fault detection and fault tolerance features. For example, running identical calculations multiple times and checking the consistency of the outcome.

Whichever type of redundancy is had been used, the costs will rise. When choosing hardware redundancy extra parts are required, also the power needs, the size of the device, and the cost of the development will increase parallel [10]. If we using software redundancy, it also has some effect to hardware redundancy, because we need stronger processor and bigger memory capacity, more time in developing phase, and so on. [11]

III. SOFTWARE REDUNDANCY

Reliability can be increased by increasing the number of the software segments. It need to be compare – with the proper algorithm – these redundant software segments, and calculate/chose the right result as a local output. This result will be one of the input parameter of the next software module. [12]

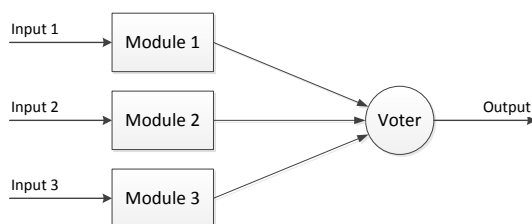


Figure 3. Triple modular redundancy

A. Majority voters

Majority voters need to have at least three different inputs. If two of the inputs are working properly, the voter will give a correct result. So, the method can tolerate only one malfunction. If more than one of the inputs gets a false signal, the output of the majority voter will be incorrect (see at Fig. 3.).

In software, there could run three different software method in parallel. The results of these software methods need to be compared by the majority voter. [13]

Majority voters are simple modules both is software and hardware realization. Therefore, it has a fairly high reliability compared with the other system modules. But, if the voter gets out of order – single point of failure – the whole systems operation becomes impossible. A solution could be, if the voters are tripled, as showed in Fig. 4. By converting functions as a sequence of sequential steps, and by incorporating voters between each level, the reliability of the system can be increased significantly [14]. This way it is possible to stop the error near to the appearance of the error, so it will not spread out to the other parts of the system. Thanks to this, the system can tolerate multiple errors if, they are appearing in different levels.

B. Software solutions

The advantage of the software solution – compared to the hardware solution – is the flexibility, less parts demand (against with a 32-bit long hardware voter), resulting in lower consumption and cost. Although, the algorithm requires only a small computing capacity, but it is slower than the dedicated hardware, and in addition, in

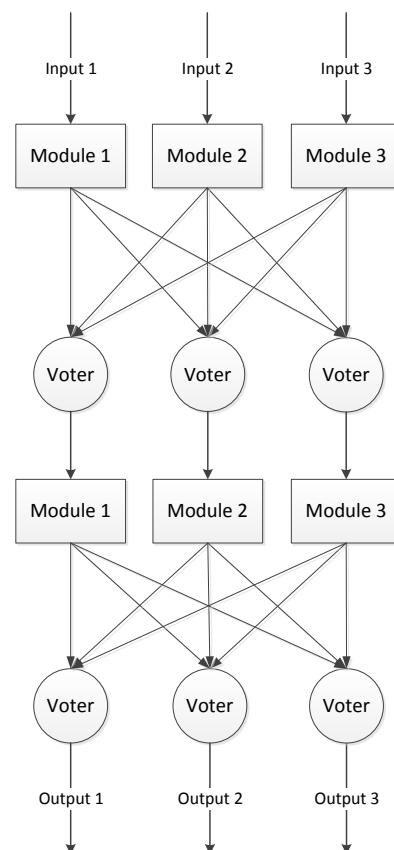


Figure 4. Sequential triple modular redundancy

the case of independent systems, synchronization problems need to be solved. [15]

C. Redundant measurements

However, many times (such as outputs of sensors), the values are correct, but they are not exactly the same, they differ within their accuracy. In a software solution makes it easy to produce the correct output. [16] A delta deviation is allowed for the measured values. If the measured values are within the delta range (relative to each other), it does not count as an error. But, in the case of multiple parallel voters, it is necessary to ensure that each of the voting outputs is exactly the same (bits are exactly the same).

If delta tolerance is selected according to the powers of two, then this method can be used for both hardware and software voters if the LSB bits are omitted from the comparison – with masking or shifting right the measured values. [17]

Mean value voter gives a different solution for the above-mentioned problem. It is providing the best result for multiple – even with significantly different – inputs. [18] As shown in Figure 5., the voter is selecting the middle value. As long as, two signs out of three are correct, the voter always choose the correct signal. The principle can be applied to any voter with an odd number of inputs.

In some cases, it makes sense to determine the output value as a function of the input values. For example, if not the middle value had been chosen – like in the mean value voter – but calculates the currently expected output signal based on the values of the inputs recorded at previous times. [19]

IV. REALIZATION

If we use only one actuator, we cannot duplicate the voters. Therefore, if possible and the nature of the process permits, several interveners and a sufficient number of voters should be used.

In our solution three digital temperature sensors output values are used as the input signals of the system, three voters had been applied, and three fan used as actuators. It is shown in Fig. 6 and Fig. 7. Fig. 8. shows the flowchart of the demonstration project.

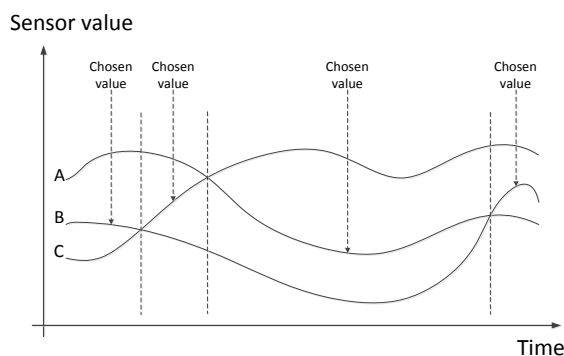


Figure 5. Technological voter

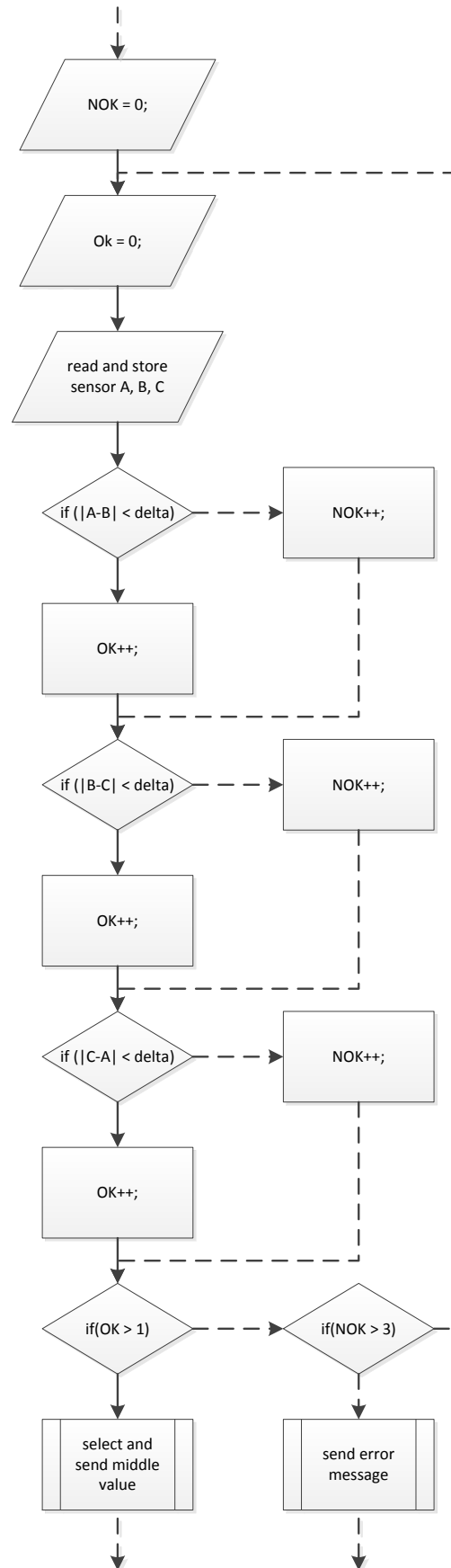


Figure 8. Flowchart of multiple majority voters

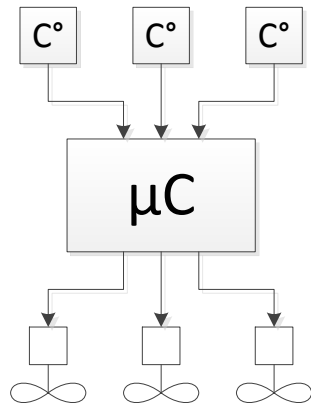


Figure 6. Operating plan of the multiple majority voter

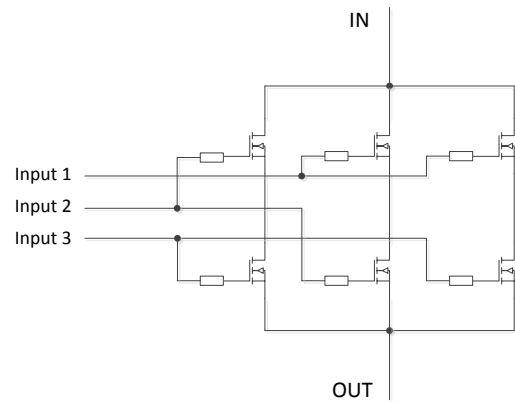


Figure 9. Technological voter

Another solution – to show a robust solution – is if the voting circuit had been left, and the interconnected system of several interveners are used, which also performs the voting task, in addition to the process control. This is called as a technology voter (Fig. 9). The voting takes place by serial parallel coupling of six FET's. Technology voting does not only have the advantages of increasing reliability due to the lack of a voting circuit, but it can also be used to replicate the interveners and to deal with errors in them.

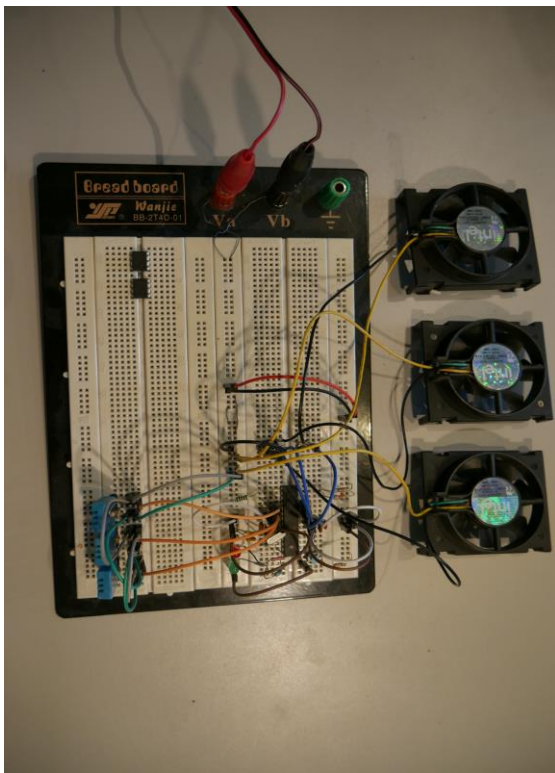


Figure 7. Realization of the multiple majority voter

V. CONCLUSION

In this paper, is showed the theoretical base connection between fault, error and failure, the possible causes of faults, fault tolerant solutions and fault tolerant software solutions. In a realization, it had been shown, a redundant software block based system, with a reliable measurement algorithm. By the mentioned solutions, the error-free running time and error detecting features can be increased, as showed in the implementation of the system. We believe that the presented methods can be used in several applications.

REFERENCES

- [1] Gy. Györök. A-class amplifier with FPAA as a predictive supply voltage control. In: 9th International Symposium of Hungarian Research on Computational Intelligence and Informatics (CINTI2008), 2008. 361–368. p.
- [2] György Györök, Bertalan Beszedes. Fault tolerant power supply systems In: Orosz Gábor Tamás 11th International Symposium on Applied Informatics and Related Areas (AIS 2016). Székesfehérvár, Magyarország, Budapest: Óbudai Egyetem, 2016. pp. 68-73. (ISBN:978-615-5460-92-0)
- [3] Györök György, Beszedes Bertalan. Duplicated Control Unit Based Embedded Fault-masking Systems. In: Szakál Anikó IEEE 15th International Symposium on Intelligent Systems and Informatics : (SISY 2017) Óbudai Egyetem. Szabadka, Szerbia. 2017. pp. 1-6. (ISBN:978-1-5386-3855-2)
- [4] Bray W. Johnson. Design and Analysis of Fault-Tolerant Digital Systems. 1989. Addison-Wesley Publishing
- [5] Gy. Györök. Embedded hybrid controller with programmable analog circuit. In: Intelligent Engineering Systems (INES), 2010 14th International Conference on. IEEE, 2010.
- [6] K. Lamár, J. Neszveda. Average probability of failure of aperiodically operated devices. In: Acta Polytechnica Hungarica, 10.(8.). 2013. 153–167. p.
- [7] Gy Györök, T Orosz, M Makó, T Treiber To Achieve Circuit Robustness by Co-operation of FPAA and Embedded Microcontroller In: Szakál Anikó (szerk.) IEEE 8th International Symposium on Applied Computational Intelligence and Informatics: SACI 2013. Konferencia helye, ideje: Timisoara, Románia, 2013.05.23-2013.05.25. (IEEE) New York: IEEE, 2013. pp. 315-320. (ISBN:978-1-4673-6397-6)
- [8] Gy. Györök. The FPAA realization of analog robust electronic circuit. In: Computational Cybernetics, 2009.

- ICCC 2009. IEEE International Conference on. IEEE, 2009.
- [9] Gy Györök Embedded hybrid controller with programmable analog circuit In: Szakál A (szerk.) 14th International Conference on Intelligent Engineering Systems: Proceedings. Konferencia helye, ideje: Las Palmas, Spanyolország, 2010.05.05-2010.05.07. Budapest: IEEE Hungary Section, 2010. pp. 1-4. (ISBN:978-1-4244-7651-0)
- [10] György Györök, Bertalan Beszédés. Artificial Education Process Environment for Embedded Systems In: Orosz Gábor Tamás (szerk.) 9th International Symposium on Applied Informatics and Related Areas - AIS2014. Konferencia helye, ideje: Székesfehérvár, Magyarország, 2014.11.12 Székesfehérvár: Óbudai Egyetem, 2014. pp. 37-42. (ISBN:978-615-5460-21-0)
- [11] J. Kopják J. Kovács. Compering event-driven program models used in embedded systems. In: Automotive-Entwicklungen und Technologien. 2011. 90-95. p.
- [12] J. Kopják. Dynamic analysis of distributed control network based on event driven software gates. In: IEEE 11th International Symposium on Intelligent Systems and Informatics. Subotica. Serbia. 2013. ISBN: 978-1-4673-4751-8. 293–297. p.
- [13] J. Kopják, J. Kovács. Implementation of event driven software gates for combinational logic networks. In: IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics. Subotica, Serbia. 2012. ISBN: 978-1-4673-4751-8. 299–304 p.
- [14] Gy Györök, M Seebauer, T Orosz, M Makó, A Selmeci Multiprocessor Application in Embedded Control System In: Szakál A (szerk.) 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, SISY 2012, Subotica, 2012, September, 20-22. Konferencia helye, ideje: Subotica, Szerbia, 2012.09.20-2012.09.22. Piscataway: IEEE, 2012. pp. 305-309. (ISBN:978-1-4673-4751-8)
- [15] Gy. Györök, M. Makó. Configuration of EEG input-unit by electric circuit evolution. In: 9th International Conference on Intelligent Engineering Systems (INES2005), 2005. 1–7. p.
- [16] Gy. Györök, M. Makó, J. Lakner. Combinatorics at electronic circuit realization in FPAA. In: Acta Polytechnica Hungarica, Journal of Applied Sciences, 2009. 6(1). 151–160. p.
- [17] Gy. Györök. The function-controlled input for the IN CIRCUIT equipment. In: 8th Intelligent, Engineering Systems Conference(INES2004), 2006. 443–446. p.
- [18] Gy. Györök. Self configuration analog circuit by FPAA. In: 4th Slovakiien – Hungariien Joint Symposium on Applied Machine Intelligence (SAMI2006), 2006. 34–37. p.
- [19] Gy. Györök, L. Vokorokos, L. Hluchý. Crossbar network for automatic analog circuit synthesis. In: 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI 2014). IEEE Computational Intelligence Society. Szerk.: J. Fodor. Budapest. 2014. ISBN:978-1-4799-3441-6, 263–267. p.

The usage of programmable logics in the education of Digital Technology

Nikoletta Tolner and András Dávid

*Campus of Alba Regia Technical Faculty of Óbuda University
H-8000 Hungary Szekesfehervar, Budai str. 45, Hungary
tolner.nikoletta@amk.uni-obuda.hu
david.andras@amk.uni-obuda.hu*

Abstract: *The authors have been teaching Digital Technology to full time and correspondent students for several years. In recent years, the manufacturers provided newer and more advanced solutions. At present, the users can get around the physical building and actual checking of the circuitry, as appearance of programmable logics opened new possibilities for circuit design and testing. The application of these novelties in the education might facilitate both the practical and theoretical knowledge of the students. In the present paper, in parallel with an electronic university lecture learning material, the process of using programmable logics in this special context was described. The XILINX's free downloadable ISE WebPACK software was used. Digital circuitries were edited with circuit diagram editor or VHDL instructions, and the operation of the circuit was simulated with ISE WebPACK software.*

I. CURRICULAR SUBJECTS

A) Digital Technology

Digital Technology is a basic subject for students in electrical engineering, informatics engineering and technical management as part of the core material. Students in electrical engineering study Digital Technology I. in the first semester for 4 credits with 2 lectures and 1 lab practice per week, Digital Technology II in the second semester for 3 credits with 2 lectures per week, then Digital Technology III with 2 lab practices per week.

Students in informatics engineering study the subject as Digital Technology with 2 lectures and 2 lab practices for 4 credits in the second semester.

Students in technical management study the subject as Analog and Digital Technology in the fourth semester with 2 lectures and 2 lab practices for 5 credits.

The aim of the subject is to achieve all the basic hardware knowledge required to design digital circuits. To this end, students deepen their theoretical knowledge through practical tasks. As lecturers, our goal is to help the students to acquire the theoretical knowledge, and to learn to use it through practical examples, like the gradual implementation of the ISE WebPACK program from XILINX environment.

B) Design of Digital Systems

Students in electrical engineering on Hardware specialization can take this course in the sixth semester. The course consists of 4 lectures and 3 lab practices for 8 credits. The goal of the course is to familiarize the students with the building blocks of digital sys-

tems, their uses, connections and diagnostic possibilities. The design possibilities of modern circuits, the basics of programmable logics, the analyzation of possible starting points of specific tasks, and design considerations are used. During laboratory practice, students use and measure the possible solutions. The main core of the course material is the detailed description of programmable logical circuits.

II. LABORATORY PRACTICE

A) Digital Technology lab

During laboratory practice, the students study the operation of sequential and combinational logic circuits; the majority of the practice is the analyzation of sequential logic circuits. Xilinx's programmable CPLD circuit (XC9572XL) (Figure 1.) was used built into a specifically designed box (Figure 2). The signals can be measured on the pins, and can be monitored with an oscilloscope. The students were capable of carrying out tasks without the need of knowing this specific programmable circuit.

Unfortunately, the following problems have arisen with this device:

- broken cables
- welding problems
- issues from multiple reprogramming

The replacement of the programmable ICs solved some of the problems for a short period of time, but has not meant permanent solution.

B) Design of Digital Systems lab

During these laboratory practises, students use the same device as in the Digital Technology lab. However, in this specific course, students not only analyse the operation of programmed circuits, but create programs as well. Using the Xilinx's ISE WebPACK circuit diagram designer, students create various circuits. Then the complete circuits are uploaded into the programmable ICs and the signals with an oscilloscope are measured. During the seminar, students get a complex task that has to be evaluated at the end of the term. The multiple uploads, unfortunately, also caused malfunctions in the ICs.

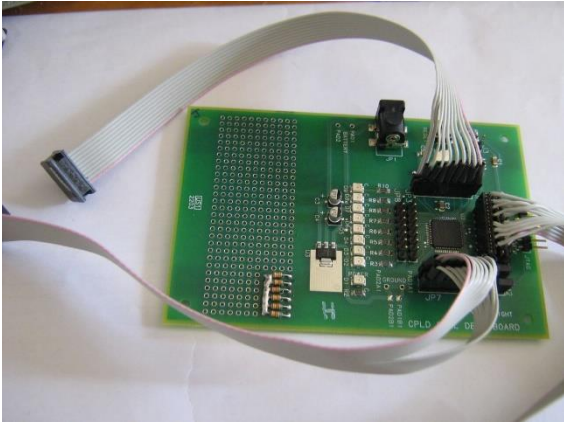


Figure 1. A board with programmable circuit



Figure 2. A measuring device

III. GOALS

In both subjects, the major goal was to substitute the classical measuring with simulations. The Xilinx ISE WebPACK has a simulation module that can be used for this purpose. However, the usage of the simulation module requires a certain level of VHDL knowledge that was integrated into the Digital Technology and Design of Digital Systems labs.

In the case of Design of Digital Systems subject, an electronic course material was also created by the Authors, focusing on the VHDL. The students work with the circuit diagram designer and the VHDL editor.

In the Digital Technology labs students analyse – with the help of the simulation module – sequential and combinational logic circuits designed in VHDL. In this subjects, only the basics of VHDL are delivered, to use the simulation module. The oscilloscope measuring still remains part of the laboratory practises besides the simulation.

IV. THE XC9500 FAMILY

This can be considered as the standard type of the Xilinx's CPLDs [1]. The components of the XC9500 family [2] essentially work in 5V TTL and CMOS systems, but can be configured to provide and receive 3.3V signals. The highest system frequency of the

device is 100MHz, and the pin-to-pin minimum delay is 5ns. The XL types requires 3.3V power supply. The input buffers fully accept 5V signals; the 3.3V output signals in 5V systems correspond with the logical "H" level that makes it possible to use it in TTL systems without additional circuits. With the proper configuration of the I/O ports the device is capable of working in 2.5V environments. In this case the pin-to-pin delay is still 5ns, but the system frequency can reach up to 222MHz. The XC9500/XL/VL family is industry's first CPLD device that builds upon a 5V FLASH memory.

The XC9572/XL's [3] features:

- 5 ns pin-to-pin logic delays
- System frequency up to 178MHz
- 72 macro cells with 1600 usable gates
- Available is small footprint packages
 - 44 pin PLCC (34 user I/O pins)
 - 44 pin VQFP (34 user I/O pins)
 - 48 pin CSP (38 user I/O pins)
 - 64 pin VQFP (52 user I/O pins)
 - 100 pin TQFP (72 user I/O pins)
- Optimized for high-performance 3.3V systems
 - Low power operation
 - 5V tolerant I/O pins accept 5V, 3.3V, and 2.5V signals
 - 3.3V or 2.5V output capability
 - Advanced 0.35 micron feature size CMOS Fast FLASH™ technology
- Advanced system features
 - In-system programmable
 - Superior pin-locking and routability with Fast CONNECT™ II switch matrix
 - Extra wide 54-input Function Blocks
 - Up to 90 product-terms per microcell with individual product-term allocation
 - Local clock inversion with three global and on product-term clocks
 - Individual output enable per output pin
 - Bus-hold circuitry on all user pin inputs
 - Full IEEE Standard 1149.1 boundary-scan (JTAG)
- Fast concurrent programming
- Slew rate control on individual outputs
- Enhanced data security features
- Excellent quality and reliability
 - Endurance exceeding 10000 program/erase cycles
 - 20 years data retention
 - ESD protection exceeding 2000V

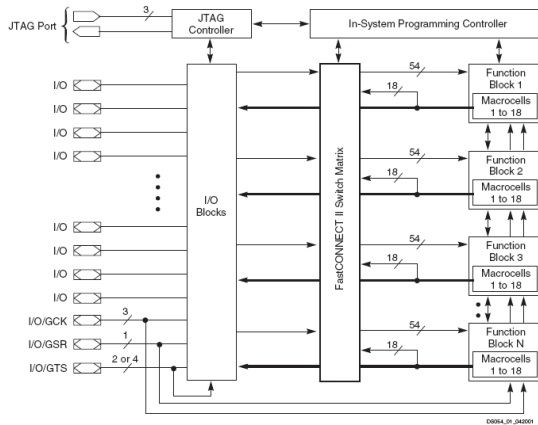


Figure 3. Structure of the XC9500XL

V. XILINX ISE WEBPACK

The Xilinx ISE WebPACK (Integrated Software Environment) is a free software developed by Xilinx for their FPGAs and CPLDs [4]. It can be downloaded from the company’s website [5]. The software contains every tool needed for circuit diagram or hardware description language based development. Digital circuits can be created in the development environment as circuit diagram (Schematic) with the help of the circuit diagram designer; or can be written in hardware description language with the HDL editor. Supported languages: Verilog and VHDL.

A) Circuit diagram based development

The goal is to make the circuit in the Xilinx ISE WebPACK based on the circuit diagram, and to analyse it with simulation, with the help of a Schematic file (circuit diagram) and a test bench file. The Schematic file contains the circuit diagram; the test bench file is necessary to run the simulation. After the installation, the development environment can be started with the ISE Design desktop icon. After starting the program, the orienting window appears (Figure 4).

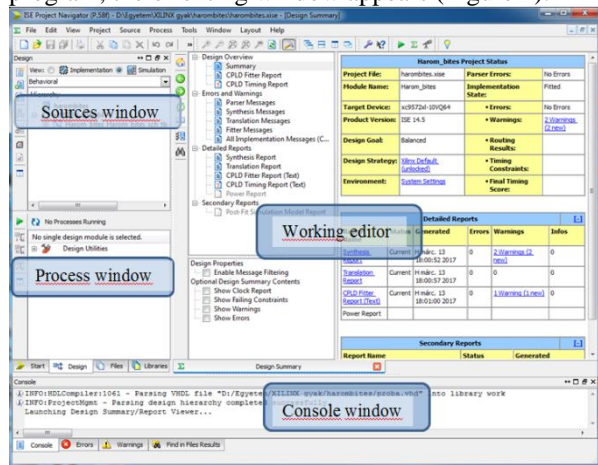


Figure 4. The program’s starting window

The development environment organizes files into projects. The first step is to create a new project. We

need to give the project a name and select a working directory, and the description of the project is also possible. In order to achieve a top level source, a circuit diagram based source (Schematic) has to be defined. In the next step, the tool type and attributes have to be defined. With this step, an empty project has been created. Then a new source file to the project is necessary. In the end, a circuit diagram designer window appears, and followed by the creation of the circuit diagram (Figure 5.). The symbols in order to create the circuit can be found on the Symbols panel (organized into categories). The user can also create symbols. After placing the symbols, they can be connected by clicking the pins. It’s a help with complex circuit diagrams that pins don’t need to be connected, it’s enough to name the cables. Pins on cables with the same name are considered to be connected in reality. After creating the circuit diagram, I/O markers needs to be defined. The program automatically names the markers, but it’s a good practice to rename them. If we receive a signal from outside port, or send a signal out, then Buffers needs to be installed between the I/O markers and the circuit.

After finishing with the circuit diagram, it’s a good practice to run a syntax check. If it finishes without error, then we can compile the Schematic file into VHDL source code to simulate.

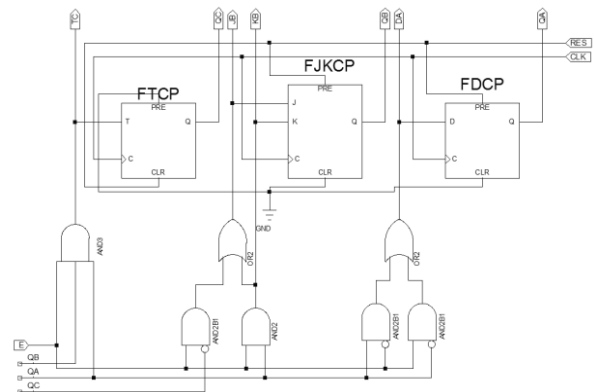


Figure 5. Example of circuit diagram creation

B) Hardware Description Language (HDL) based development

In order to test the circuitry within the HDL (VHDL), the source file and the test bench file is needed. The VHDL source file contains the code that describes the digital circuit; the test bench file is necessary to run the simulation.

The first step is to create a new project, and then to add a new source file. This is the same as with the circuit diagram designer, however, for a top-level source type the hardware description language based (HDL) source type needs to be selected, also the new source file type is now HDL based (VHDL Module). Adding a new source file, the New Source Wizard’s Define Module window appears, where the ports of the device have to be defined (Figure 6).

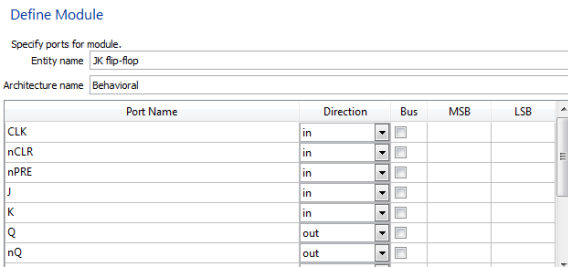


Figure 6. Definitions of the JK flip-flop's ports

After the appearance of the source editor window, the VHDL source code can be inserted (Figure 7, a source code describing a JK flip-flop).

```

library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
entity JK_ff_74LS76A is
    Generic (tpLHCLK_Q : time := 20 ns;
            tpHLCLK_Q : time := 20 ns;
            tpLHnPRE_nCLR_Q : time := 20 ns;
            tpHLnPRE_nCLR_Q : time := 20 ns;
            tsu_CLK : time := 20 ns);
    Port ( CLK : in STD_LOGIC;
          J : in STD_LOGIC;
          K : in STD_LOGIC;
          nCLR : in STD_LOGIC;
          nPRE : in STD_LOGIC;
          Q : out STD_LOGIC;
          nQ : out STD_LOGIC);
end JK_ff_74LS76A;

architecture Behavioral of JK_ff_74LS76A is
    signal J_act : STD_LOGIC := '0';
    signal K_act : STD_LOGIC := '0';
    signal next_state : STD_LOGIC := '0';
    signal not_next_state : STD_LOGIC := '1';
begin
    JK_ff : process(J, K, CLK, nCLR, nPRE)
    begin
        if NCLR = '0' then
            if next_state = '1' then
                next_state <= '0' after tpHLnPRE_nCLR_Q;
                not_next_state <= '1' after tpLHnPRE_nCLR_Q;
            end if;
        elsif nPRE = '0' then
            if next_state = '0' then
                next_state <= '1' after tpLHnPRE_nCLR_Q;
                not_next_state <= '0' after tpHLnPRE_nCLR_Q;
            end if;
        elsif J /= J_act or K /= K_act then
            J_act <= J after tsu_CLK;
            K_act <= K after tsu_CLK;
        elsif rising_edge(CLK) then
            if J_act = '1' and K_act = '0'
               and next_state = '0' then
                next_state <= not(next_state) after tpLHCLK_Q;
                not_next_state <= not(not_next_state)
                    after tpHLCLK_Q;
            elsif J_act = '0' and K_act = '1'
               and next_state = '1' then
                next_state <= not(next_state) after tpHLCLK_Q;
                not_next_state <= not(not_next_state)
                    after tpLHCLK_Q;
            elsif J_act = '1' and K_act = '1' then
                if next_state = '0' then
                    next_state <= not(next_state) after tpLHCLK_Q;
                    not_next_state <= not(not_next_state)
                        after tpHLCLK_Q;
                else
                    next_state <= not(next_state) after tpHLCLK_Q;
                    not_next_state <= not(not_next_state)
                        after tpLHCLK_Q;
                end if;
            end if;
        end if;
    end process JK_ff;
    Q <= next_state;
    nQ <= not_next_state;
end Behavioral;
    
```

Figure 7. Source code of a JK flip-flop

After writing the VHDL source code, the check of syntax and the compilation is taken place.

C) Simulation

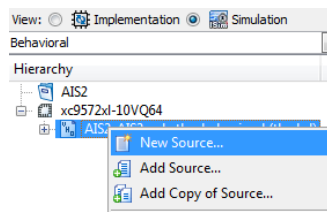


Figure 8. Adding new source to the project

The operation of logic circuits that has been created either as circuit diagram or with hardware description language can be tested with the ISIM simulator. In order to achieve a simulation

to achieve a simulation, a test bench file has to be created, and then added to the project. The test bench describes the signals sent to each input. If the circuit contains clock signal, then a clock signal generator processor has to be added to the test bench file (Figure 9).

```

-- Clock period definitions
constant CLK_period : time := 10 ns;
-- Clock process definitions
CLK_process : process
begin
    CLK <= '0';
    wait for CLK_period/2;
    CLK <= '1';
    wait for CLK_period/2;
end process;
    
```

Figure 9. Definition of clock signal

The input signals definitions should be written in the appearing editor (Figure 10).

```

-- *** Test Bench - User Defined Section ***
tb : PROCESS
BEGIN
    E <= '1';
    RES <= '0'; wait for clk_period;
    RES <= '1'; wait for clk_period; RES <= '0';
    wait for 2*clk_period; E <= '0';
    wait for clk_period; E <= '1';
    WAIT; -- will wait forever
END PROCESS;
-- *** End Test Bench - User Defined Section ***
    
```

Figure 10. Enter excitation of input signals

Before running the simulation, the setup of the simulation process properties is required. This can be done within the menu shown on Figure 11. The Simulation Running Time is probably the most important property.

After the clock signal and the input signal setup is done, the ISIM simulator can be started. The most important part of this is the waveform windows, where the timing diagram of the circuit can be seen (Figure 12.). The timing diagram describes the operation of the circuit shown on Figure 5.

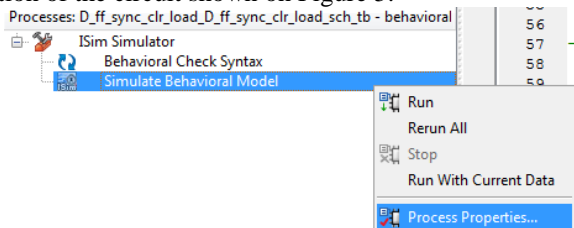


Figure 11. Setup of the properties of the simulation process

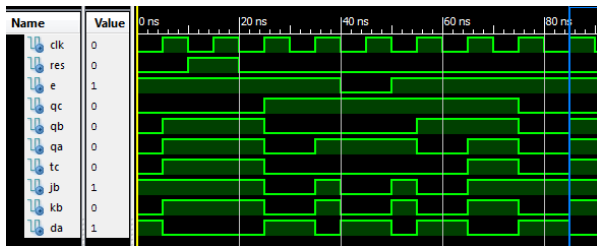


Figure 12. Timing diagram of the tested circuit

VI. ELECTRONIC COURSE BOOK

A. Goals

There are many great literature of the subjects for the students [6-8], albeit these references do not fully satisfy our criteria in the curriculum. To this order, a new course book will be made for the Design of Digital Systems subject. The book, additional to the theoretical material, would contain a good number of examples, to help deeper and practical understanding of the subject. At the labs, unfortunately there is no time to start from the basics, and there is a need for a course book that the students can use to prepare in advance. The simulation related parts of the course book also can be used in the Digital Technology labs. We aim to keep the theoretical descriptions at the necessary minimum, because we want to emphasize the practical examples.

B. The structure of the electronic course book

The course book is made up of 4 main chapters. In the first part, there are theoretical descriptions, and the second part would contain a good number of examples.

The first chapter would be a summary about programmable logics. The second chapter is the introduction of the Xilinx ISE WebPACK program suit, with a detailed description from the downloading of the program till its usage. In the third chapter would contain a description about the VHDL. An extensive material into the course book from the language elements to its syntax has been incorporated. Here only some examples and program parts can be found. In the fourth chapter, examples of combinational and sequential logics are described. For every examples, only a short description would be addressed, because those enrolling the course already possess a certain level of Digital Technology knowledge (the successful completion of the Digital Technology course is a requirement of this course). After the description of the examples, the description of VHDL would occur the test bench file for the simulation, and the simulation result timing diagram. These examples will help the students to solve more complex tasks later. In most cases, the whole test bench file is provided in order to use them in Digital Technology labs. In the fourth chapter, we added exercises after each topic that we would like to give as homework to the Design of Digital Systems course's students. We discuss the solutions at the laboratory practices. Later we plan to make the exer-

cises' solutions available to the students in electronic format. Figure 13. shows a part of the course book.

```

LIBRARY ieee;
USE ieee.std_logic_1164.ALL;

ENTITY tb IS
END tb;

ARCHITECTURE behavior OF tb IS
    -- példaművelet
    COMPONENT Felossz_2
    PORT (
        A : IN std_logic;
        B : IN std_logic;
        S : OUT std_logic;
        C : OUT std_logic
    );
END COMPONENT;

--Inputs
signal A : std_logic := '0';
signal B : std_logic := '0';

--Outputs
signal S : std_logic;
signal C : std_logic;

BEGIN
    -- be-, kimenetek megfeleltetése
    uut: Felossz_2 PORT MAP (
        A => A,
        B => B,
        S => S,
        C => C
    );

    -- Stimulus process
    stim_proc: process
    begin
        A <= '0'; B <= '0'; wait for 20 ns;
        A <= '0'; B <= '1'; wait for 20 ns;
        A <= '1'; B <= '0'; wait for 20 ns;
        A <= '1'; B <= '1'; wait for 20 ns;
        A <= '0'; B <= '0'; wait for 20 ns;
        A <= '0'; B <= '1'; wait for 20 ns;
        A <= '1'; B <= '0'; wait for 20 ns;
        A <= '1'; B <= '1'; wait for 20 ns;
        wait;
    end process;
END;

```



Figure 13. Part of the Design of Digital Systems course book

C. Expectations, hopes

We hope that our course book will help the students to gain the necessary knowledge in the subject. Of course it would be impossible to include everything into one course book, but we try to demonstrate the basics with a great number of examples. Our goal always was to help the understanding with as many practical examples as possible. This electronic course book was also made with the same aim in mind. We hope that it will live up to the expectations. Our future plan is to create an online course in the subject.

VII. REFERENCES

- [1] <https://www.xilinx.com/cpld/>
- [2] https://www.xilinx.com/support/documentation/data_sheets/DS063.pdf
- [3] https://www.xilinx.com/support/documentation/data_sheets/ds057.pdf
- [4] <https://www.xilinx.com/products/design-tools/ise-design-suite/ise-webpack.html>
- [5] <https://www.xilinx.com/support/download/index.html/content/xilinx/en/downloadNav/design-tools/archive.html>
- [6] László Varga: FPGA development in Xilinx ISE Webpack <https://www.hobbielektronika.hu/forum/getfile.php?id=118958>
- [7] Gergely Nagy, Péter Horváth, András Poppe: Practical aspects of thermal transient testing in live digital circuits (Proceedings of the 19th International Workshop on THERMAL INvestigation of ICs and Systems, ISBN:978-1-4799-2271-0) <https://repozitorium.omikk.bme.hu/handle/10890/5088>
- [8] Szabolcs Szilágyi: Multiway Switching Controller Design using FPGA (ICAI 2014: Proceedings of the 9th International Conference on Applied Informatics) <http://icai.ektf.hu/icai2014/papers/ICAI.9.2014.2.57.pdf>

THE CLASSIFICATION OF THE APPROACH OF CAVE ENTRANCES

Peter Tarsoly, PhD.

Alba Regia Technical Faculty of Óbuda University, Székesfehérvár, Hungary

tarsoly.peter@amk.uni-obuda.hu

Abstract - From the viewpoint of any research the most essential characteristic of the caves are that how easily are they passable, what kind of technical equipment and preparedness need to being in service in a cave. The existence of the map data and logistic information are the fundamental information. From this viewpoint the approach of the entrances has only a secondary role. If a cave research needs a moving of a lot or a heavy equipment, the approach of a cave entrance is not negligible. The numerical characterization of topographical and logistical parameters means an easy and quick solution for this problem.

I imagined the index-number (K), which expressed the difficulty of the approach of entrances as the sum of four parameters: distance from a beaten road (Lh), elevation above a beaten road (Lm), steepness in degrees (M), vegetation covering (N). By the calculation of the value K, I take into account the Lm and M values with twofold weights, because these two numbers are the most important terms of the judgement of the difficulty level.

For testing the classification parameters I selected 39 caves, cave-like objects, tallus caves, artificial caves and rocky shelters in the Velence Hills (Pákozd, Sukoró, Pátka, Lovasberény, Pázmánd).

I compared the difficulty level of the approach of caves, which are in granite (Granite rocks of the Meleg-hegy Natural Conservation Area, Rocking stones of Pákozd Natural Conservation Area) and in andesite (Quartzite rocks of Pázmánd Natural Conservation Area) in the Velence Hills.

The difficulty level of the approach of granite caves is easy – medium (K=7), and by andesite caves medium – troublesome (K=10). Neither at granite caves nor at andesite caves the distance from a beaten road is not considerable (Lh<50 m). The elevation value above a beaten road is small (Lm=0-10 m) by both cave types. The steepness is medium by both cave types (M=20-40°). The vegetation covering is not significant by granite caves, but is very thick by andesite caves.

I. INTRODUCTION

Cave entrances constitute the part of the space which can be written down with topographical elements. They are punctiform objects, which are shown on maps with the Greek omega (Ω) letter, as a symbol. The symbols are often generalized, pushed away from their real place, if map uncovering and plane drawing contents make this necessary. From the map elements which can

be found in the environment of the entrances or from the colour usage fundamental information can be read onto the environment of the entrance relevantly – type of the vegetation, steepness of the terrain based on the contour lines, routes, glades and paths near the cave etc. Onto the detailed analysis of the approach of cave entrance these information are not sufficient.

From the viewpoint of any research the most essential characteristic of the caves are that how easily are they passable, what kind of technical equipment and preparedness need to being in service in a cave. The existence of the map data and logistic information are the fundamental information. From this viewpoint the approach of the entrances has only a secondary role. If a cave research needs a moving of a lot or a heavy equipment, the approach of a cave entrance is not negligible (Figure 1.).



Figure 1. Transporting the Leica C10 Scanstation in the Kőmosó-árok near Csesznek, in the Bakony Mountain

Onto the case of granite caves I drew up the content and formal requirements of the cave entrance thematic map [1]. The cave entrance thematic maps make possible to analyse the environment and approach of cave entrances from the viewpoint of topography and logistic; but their making is time-consuming and complicated in some cases. The numerical characterization of topographical and logistical parameters means an easy and quick solution.

II. THE PREDICAMENTAL VIEWPOINTS OF THE APPROACH OF THE CAVE ENTRANCES

Based on numerous field researches it is possible to develop a uniform system which can be applied to all caves. The predicamental system which was sketched by

me can be considered only for an initial solution of a long-term examination. With the increase of the number of the caves involved in the examination the predicamental viewpoints can be refined. Other viewpoints may occur and may be it should be necessary to relocate the weights for the determination of the definitive index-number.

I imagined the index-number (K), which expressed the difficulty of the approach of entrances as the sum of four parameters: distance from a beaten road (L_h), elevation above a beaten road (L_m), steepness in degrees (M), vegetation covering (N).

$$K = L_h + L_m + M + N \quad (1)$$

By the calculation of the value K , I take into account the L_m and M values with twofold weights, because these two numbers are the most important terms of the judgement of the difficulty level. This latter two viewpoints are immutable, since while it is possible to extend the road to the cave and it is possible to thin out the vegetation with mechanical or chemical devices; steepness and elevation are constant parameters (Figure 2.).



Figure 2. The entrance of some caves opens on a very steep place (Polák-hegyi-álbarlang)

For the calculation of the single parameters I used numerical values uniformly from zero to four. It makes possible the comparison between them. The Table 1. contains the parameters belonging to the index-numbers.

Table 1. The classification of the parameters, which are expressing the difficulty of the approach of a cave entrance

	L_h [m]	L_m [m]	M [°]	N	K
0	It is possible to approach the cave on a road.	The entrance is in a level with the road.	0	none	0-8 point, easy
1	0-50	0-10	0-20	rare	8-16 point, medium
2	50-100	10-30	20-40	medium	
3	100-200	30-50	40-60	thick	16-24 point, troublesome
4	200 -	50 -	60 -	very thick	

The judgement of the vegetation covering is based on subjective viewpoints and characterize the entrance only in a given year. It is necessary to measure the steepness with more occasions between the road and the entrance, and the final parameter will be the arithmetic mean of the measured values. If it is possible, to approach the cave on a road, the steepness must be measure on the last 10 meters stretch of the road, only once a time.

The evaluation of the distance and elevation of the entrance from and over a beaten road, was based on the estimation of the step number. It should be possible to use a pedometer or an odometer, but I used a plainer solution. The normal stride (S_n) of all men exists as the function of his height. For my own height I defined this so, that I walked along a 10 meters horizontal section ten times, and averaged the results. My normal stride for horizontal terrain and comfortable walking is 75 centimetres. The normal stride varies in the function of the steepness, and it is important to know exactly this variations, because the estimation of the distance of a cave entrance from a beaten road depends on them [2]. The height of a stride depends on the length of a stride and on the steepness of the terrain. I determined the height of a stride with simple trigonometrical functions, using the length of a stride and the elevation angle (steepness). The Table 2. shows the length and height of a stride for different terrains.

Table 2. The variation of the length and height of a stride

Steepness [°]	Length of a stride [cm]				Height of a stride [cm]	
	upward	downward	upward	downward	upward	downward
0	S_n	S_n	75	75	0	0
5	$0.91 \cdot S_n$	$0.97 \cdot S_n$	68	73	6	-6
10	$0.81 \cdot S_n$	$0.94 \cdot S_n$	61	71	10	-12
15	$0.73 \cdot S_n$	$0.91 \cdot S_n$	55	68	15	-18
20	$0.65 \cdot S_n$	$0.87 \cdot S_n$	49	65	18	-24
25	$0.58 \cdot S_n$	$0.78 \cdot S_n$	44	59	20	-27
30	$0.49 \cdot S_n$	$0.65 \cdot S_n$	37	49	21	-28

Over 30°degrees steepness it is necessary to use the value calculated for 30°degrees steepness, except vertical walls. In this occasion the horizontal distance is zero, the elevation above the road must be measure with

some other methods, for example estimation of the length of a rope, trigonometrical height measurements etc. The calculation of the distance from a beaten road and the elevation above the road is based on the values in the Table 2. The result will be in meters, if we multiply the value (belonging to a known steepness) in centimetres with the number of strides, and then divide it with one hundred.

$$L_{h[m]} = \frac{\text{length of a stride}_{[cm]}^{f(\text{steepness})} \cdot \text{number of strides}}{100}$$

and

$$L_{m[m]} = \frac{\text{height of a stride}_{[cm]}^{f(\text{steepness})} \cdot \text{number of strides}}{100} \quad (2)$$

III. THE CAVES USED FOR THE EXAMINATION

For testing the classification parameters I selected 39 caves, cave-like objects, tallus caves, artificial caves and rocky shelters in the Velence Hills (Pákozd, Sukoró, Pátka, Lovasberény, Pázmánd). The Table 3. contains the cadastral number and regional position of the caves involved in the examination.

Table 3. Caves in the examination

Velence Hills			
Rocking stones of Pákozd Natural Conservation Area			
Zsivány-barlang	4510-2	Osztott-barlang	4510-512
Gömb-kő barlangja	4510-503	Teraszos-barlang	4510-515
Háromszájú-barlang	4510-504	Gomba-kő barlangja	4510-516
Iker-kő barlangja	4510-505	Rejtekek-barlang	4510-519
Kis-barlang	4510-507	Mohás-barlang	4510-524
Oroszlán-kő barlangja	4510-511	Siklóbőrös-sziklaeresz	4510-533
Granite rocks of the Meleg-hegy Natural Conservation Area			
Bárcaházi-barlang	4510-501	Páfrányos-barlang	4510-528
Likas-kő	4510-509	Kőrözsa-álbarlang	4510-529
Polák-hegyi-álbarlang	4510-525	Cserkupacos-barlang	4510-532
Borjú-völgyi-álbarlang	4510-518	Diétás-barlang	4510-534
Róka-lyuk-barlang	4510-520	Cserepes-barlang	4510-535
Bujdosó-barlang	4510-521	Tiborc-völgyi-álbarlang	4510-536
Pókhálós-barlang	4510-522	Tiborc-völgyi-átjáróbarlang	4510-537
Szúnyogos-barlang	4510-523		
Quartzite rocks of Pázmánd Natural Conservation Area			
Hasadék-barlang	4510-1	Szedres-barlang	4510-514
Pirofillit-bánya barlangja	4510-3	Kökényes-barlang	4510-517
Endrina-barlang	4510-502	Kuszoda-álbarlang	4510-526
Lapos-barlang	4510-508	Pázmándi-sziklakapu	4510-527
Maléza-barlang	4510-510	Csúzli-álbarlang	4510-530
Pilléres-barlang	4510-513	Gyümölcsöző-álbarlang	4510-531

IV. RESULTS

Figure 3-5. are showing the parameters, which are describing the approach of the cave entrances. All parameters were rounding up according to the mathematical rules, and are showing an averaged value for the single natural conservation areas.

The approach of the caves in the Granite rocks of the Meleg-hegy Natural Conservation Area (Figure 3.) is easy or medium (K=9). The distance from a beaten road is not considerable (Lh<50 m), the value of the elevation above a beaten road is medium (Lm=10-30 m), and the steepness is also medium (M=20-40°). Vegetation does not cover the entrance of the caves.

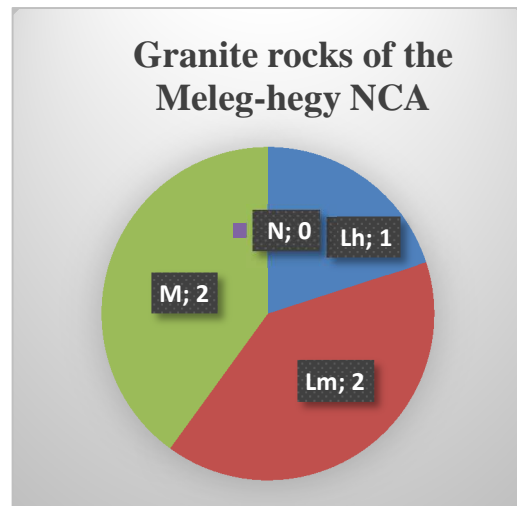


Figure 3. The difficulty level of the approach of the caves in the Granite rocks of the Meleg-hegy Natural Conservation Area

The approach of the caves in the Rocking stones of Pákozd Natural Conservation Area (Figure 4.) is easy (K=4). The distance from a beaten road is not considerable (Lh<50 m), the value of the elevation above a beaten road and a steepness is also negligible (Lm=0 m; M=0-20°). Rare vegetation covers the entrance of the caves.

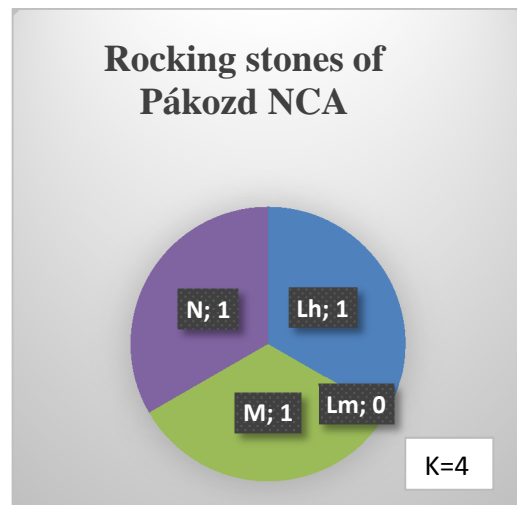


Figure 4. The difficulty level of the approach of the caves in the Rocking stones of Pákozd Natural Conservation Area

The approach of the caves in the Quartzite rocks of Pázmánd Natural Conservation Area (Figure 5.) is medium ($K=10$). The distance from a beaten road is not considerable ($L_h < 50$ m), the elevation value above a beaten road is small ($L_m = 0-10$ m), the steepness is medium ($M = 20-40^\circ$). Thick vegetation covers the entrance of the caves.

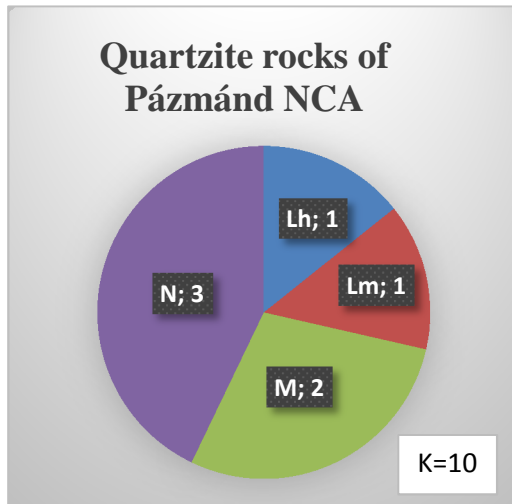


Figure 5. The difficulty level of the approach of the caves in the Quartzite rocks of Pázmánd Natural Conservation Area

Figure 3-5. give us an opportunity to compare the difficulty level of the approach of caves, which are in granite (Granite rocks of the Meleg-hegy Natural Conservation Area, Rocking stones of Pákozdi Natural Conservation Area) and in andesite (Quartzite rocks of Pázmánd Natural Conservation Area) in the Velence Hills.

The difficulty level of the approach of granite caves (Figure 6.) is easy – medium ($K=7$), and by andesite caves medium – troublesome ($K=10$).



Figure 6. A typical granite cave in the Granite rocks of the Meleg-hegy Natural Conservation Area (Cserkúpacsos-barlang)

Neither at granite caves nor at andesite caves (Figure 7.) the distance from a beaten road is not considerable ($L_h < 50$ m).

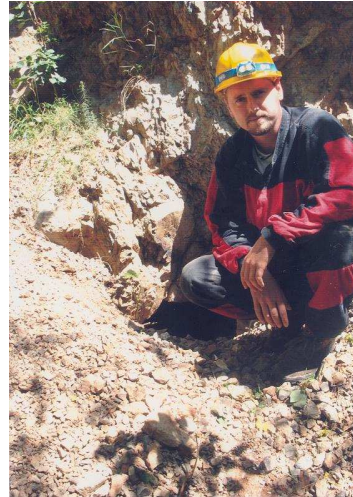


Figure 7. A typical andesite cave in the Quartzite rocks of Pázmánd Natural Conservation Area (Pirofillit-bánya barlangja)

The elevation value above a beaten road is small ($L_m = 0-10$ m) by both cave types. The steepness is medium by both cave types ($M = 20-40^\circ$). The vegetation covering is not significant by granite caves, but is very thick by andesite caves.

REFERENCES

- [1] Tarsoly P.: „A DGPS-technika pontosságának jellemzése a barlangkataszter, helyszínrajzok és térképezés pontosságának szempontjából”, MKBT Vulkanoszeleológiai Kollektívájának évkönyve, Isztimér, pp. 63-79, 2012
- [2] Berg A.: „Geographisches Wanderbuch”, Teubner Verlag, Leipzig, p. 304, 1914

CAD/CAM Systems in Implant Design in Kazakhstan

O.Y. Shvets*

* D. Serikbayev East-Kazakhstan State Technical University/Instrument Engineering, Automation and Control Sub-Department, Ust-Kamenogorsk, Kazakhstan
olga.shvets75@gmail.com

Abstract—It is considered in the work some problems and perspectives Delcam CAD/CAM systems applying in implant design and for its production on CNC machines.

I. INTRODUCTION

The development and production of medical devices are among the most intensively developing areas of scientific and technical activity. They include the development of new materials, design, production and quality control technologies. In the 21st century, medical science and technology became one of the main driving forces of modern technical civilization, gradually pushing astronautics and becoming one with information technologies, intensively developing in the last 30 years. According to the forecasts of the British government, the most popular specialists until 2030 will be bioengineers, developing new medical products, and doctors using high-tech methods of treatment [1-10].

A big problem in the dental industry today in Kazakhstan is a small number of large-scale scientific researches related to the comparative analysis of whole groups of medical products that are truly independent.

Analysis of the development of the Kazakhstan market of medical products showed that the volume of the domestic market of medical products increased most intensively from 2006 to 2012, but it is significantly inferior to the markets of leading foreign countries such as Germany, Switzerland and Austria [2, 11, 12].

The main materials used for the manufacture of implants are metal alloys (titanium, cobalt, stainless steels), polymers and ceramics. Despite the intensive increase in the use of polymers and ceramic materials in implantable products, metallic materials retain their leading role (about 60% of all implants). The share of products made of titanium alloys can be estimated at about 28% [13-15].

II. THE GOAL OF THE RESEARCH

Medical implants are implanted in the body for the purpose of prosthetics of damaged organs. As a rule, the implant has a common basic construction that can be presented in the form of a discrete size series to reduce production costs, and a unique individual part whose geometry is determined by the physical features of the structure of the patient's body.

The initial data for implant design is obtained by computer tomography of the patient. Detection and analysis of violations is done by a doctor who decides on prosthetics and together with a technologist builds a 3D

model of the implant in a specialized CAD system. The finished result of the design is the CAD model of the implant and the necessary tooling for its manufacture, as well as the results of CAE-calculation of the stress-strain state of the implant under the action of operational loads. This approach allows to ensure the production of a quality implant and positive results of prosthetics.

Implants are manufactured in various ways, including plastic deformation from sheet blanks. This is due to the fact that sheet constructions are easier and cheaper (although in some cases they may not be sufficiently hard). For the manufacture of implants, often use solid sheet or perforated (such as "mesh") blanks of titanium VT1-0, from which products with a complex spatial shape are obtained by means of plastic deformation.

The goal of the research was to develop a technology for designing and manufacturing implants of complex spatial shape and specialized modules of integrated CAD based on CAD / CAM / CAI Delcam systems designed for calculating and constructing parameterized die tooling, as well as designing the technology of its manufacturing on CNC machines.

III. DEVELOPMENT OF METHODS FOR DESIGNING AND MANUFACTURING IMPLANTS FOR MAXILLOFACIAL SURGERY

As a rule, the implant and technological equipment for its manufacture can be divided into two components: a universal and individual for each patient part, therefore the technology of computer-aided design should be designed for the production of both these parts. Tools can also be used in the process of manufacturing implants from sheet blanks on which the plastic deformation of the workpieces is carried out manually, by punching with polyurethane or in a combined method with manual finishing.

The model for the shaping of the implant blank by plastic deformation also consists of an individual patient shape part providing for the qualitative positioning and fixing of the implant on the jaw and a base with a uniform shape for installation in a container with polyurethane. The unfolding of the implant's blank should be located in the container and therefore determines its transverse dimensions.

Let's consider the process of designing an implant using a specific example. The initial results of the tomography and the shape of the implant were provided by our colleagues from the Kazakh National Medical University. The CAD model of the part of the implant in place of the lost part of the bone for osteosynthesis in the maxillofacial surgery is shown in Fig. 1.

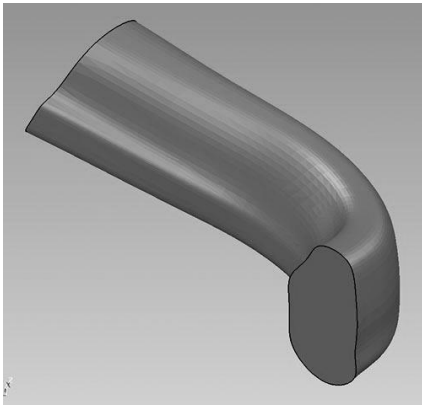


Figure 1. The CAD model

The method of constructing the surface of the implant should be universal and allows for the possibility of adjusting the shape. We developed a generalized algorithm using secant planes, which makes it possible to obtain an array of implant cross sections by planes perpendicular to its generatrix. The number of cutting planes depends on the accuracy of the design, but too many of them can lead to a decrease in computer performance. The presence of sections provides the possibility of correcting the profile of the implant with the subsequent construction of its surfaces, taking into account the shape of the jaw bones and the thickness of the sheet blank, as well as the surface of the template for its preparation. It should also be noted that the surface of the template differs from the desired shape of the implant due to the fact that in the process of spatial deformation, the titanium blank is springing.

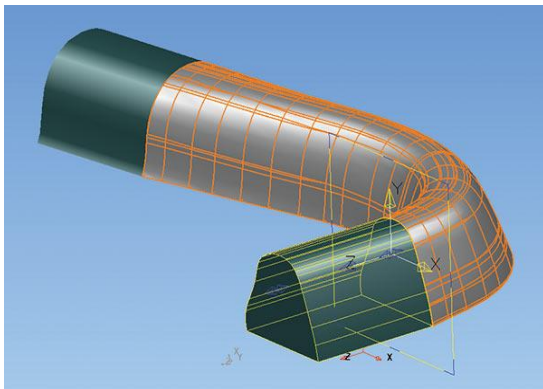


Figure 2. 3D model

Implant construction begins with the initial model obtained as a result of processing the patient's tomogram importing into the CAD system of PowerSHAPE. The resulting CAD model is a copy of the original 3D model and is suitable for further technological development. Then, on the jaw, parts of the surfaces are identified, to which the surfaces of the leaf implant will adhere when it is installed. The edges of the implant must therefore conform to the surfaces for the secure attachment of the implant to the bones of the jaw. It is necessary to add supplements both the ends along the length and the lateral generatrix of the implant to design contiguous surfaces. The first constructive supplement along the length ensures the placement of the implant over fragments of the bone for fastening with screws, and the second is needed for

designing the model to avoid overhanging the workpiece over its end edges. Subsequently, the technological allowance along the lateral generatrix of the workpiece is cut off during the finishing of the product.

At the next stage, a set (array) of cutting planes is built along the entire body of the implant, and one of the lateral generatrices of the implant base is used as the guiding curve for their construction (Fig. 3).

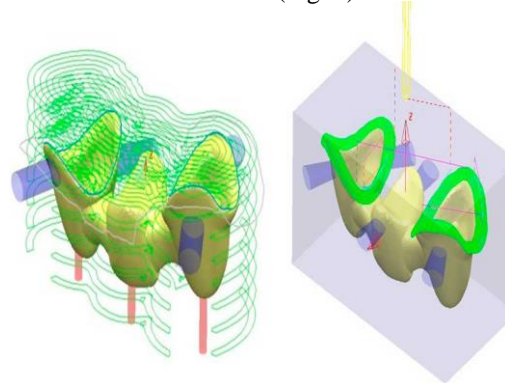


Figure 3. A set (array) of cutting planes is built along the entire body of the implant.

The surface of the implant is displaced equidistantly to take into account the thickness of the sheet. In addition to adjusting the shape of the implant, the use of cross-sections makes it possible to calculate a sweep to obtain the shape of a flat sheet blank (Fig. 4). By adjusting the array of sections of the template, it is also possible to compensate for the spring of the workpiece during the stamping process.

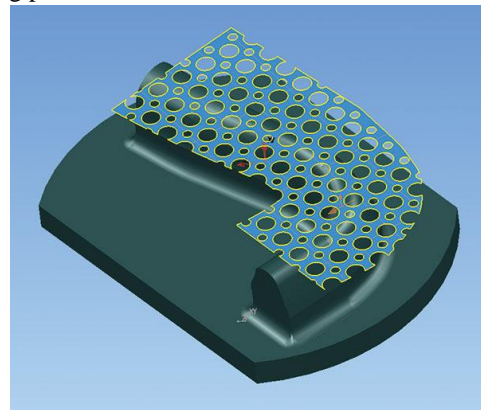


Figure 4. The surface of the implant

The process of punching a sheet blank also requires the creation of a transitional section with a round at the base on the model. As a result, we designed a ready-made model for the manufacture of a sheet implant, the base of which is designed for a universal container for the formation of a polyurethane implant.

It was modeled in the CAE system of Abaqus Student Edition to determine the loads acting on the implant during operation (during chewing). We used elements of the linear order type C3D8H contained in the standard library, from the category of 3D Stress to sample the volume and calculations that is working in all three directions in terms of their volume in a hybrid formulation. The mesh was uniform in length and thickness of the plate. Several variants of external loads have been modeled, including loads that arise when chewing both the intact and the prosthetic side of the jaw. Numerical analysis showed that the equivalent stresses

arising in the implant in the places of its fixation with screws are close to the ultimate strength of the titanium alloy. The increase in the number of points of attachment of the implant due to the addition of a second row of screws allowed to reduce the load to acceptable levels for bone and implant material.

IV. DEVELOPMENT OF CAD HARDWARE FOR THE MANUFACTURE OF IMPLANTS BY PLASTIC DEFORMATION ON THE BASIS OF THE CAD-SYSTEM POWERSHAPE

It has been developed CAD KazImplant in C # .NET language to automate the construction of stamp equipment, which is integrated with the CAD system of PowerSHAPE and performs 3D-constructions in it on the base of parametric models of stamp details. Integration is provided by using the library of API functions. It was also additionally used the specially developed library of functions, which is a superstructure over macros and implements work with objects in PowerSHAPE to build 3D models of stamp details. The main window of the developed CAD KazImplant is shown in Fig. 5.

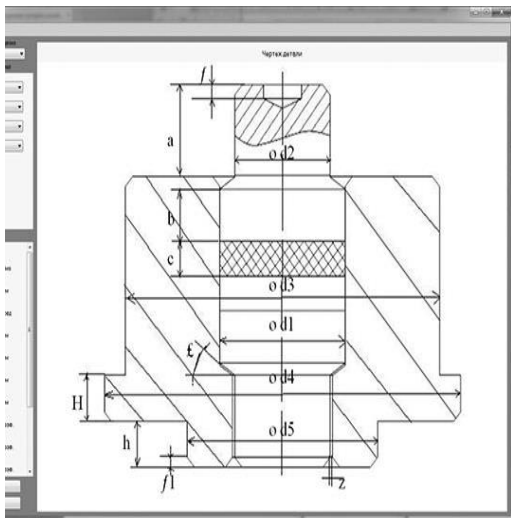


Figure 5. The main window of the developed CAD KazImplant

The parametric model of the tooling allows to arbitrarily changing the dimensions of the elements, preserving the configuration and integrity of the design by specifying the dimensional relationships (dependencies) of the variable parameters from several basic ones. The values of the basic and variable parameters can be stored in the database or set during the construction process by the user.

The details of the parts are stored in the database. The peculiarity of the application is that the user interface is dynamically formed depending on the type of product that we want to design, so it must be selected before working with the application, after that the user interface will show a drawing of the longitudinal section of the die tooling, as well as lists of parts and their parameters. Lists of details, parameters, as well as a list of types of projected products, are downloaded from the database. After setting the values of all necessary (basic) parameters and clicking the Build button, the algorithm for calculating the geometric dimensions of the parts is started, and the application builds the product in the PowerSHAPE environment in the

automatic mode. The results of the CAD KazImplant are shown in Fig. 6.

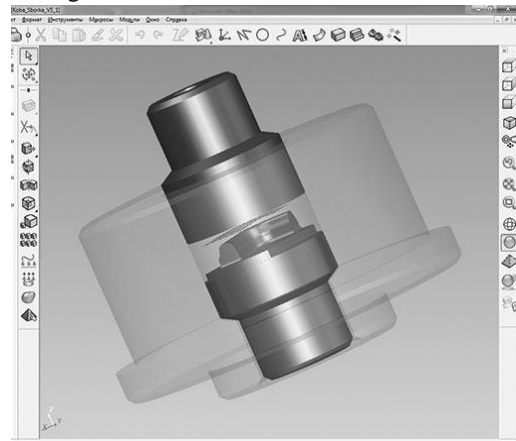


Figure 6. The results of the CAD KazImplant

We demonstrate the capabilities of the impCAD module in the example of manufacturing a designed model (a deforming tool for die tools) for the manufacture of implants.

The process of designing a product processing technology in the CAM system is based on the use of a universal visual interface and macros. The purpose of this work was to simplify the interaction of the technologist with the CAM-system and save his time in performing the same procedures by creating a specialized visual interface directly for this type of parts.

The created impCAD module can directly access the CAM API of the PowerMILL system and, with the help of macros, perform the necessary work on the development of control programs for CNC machines (Fig. 7).

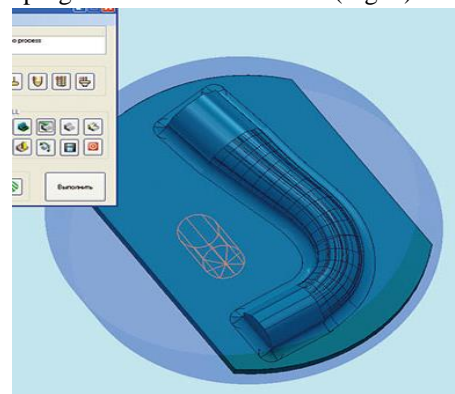


Figure 7. The results of the CAD KazImplant

The advantage of impCAD is the ability to transfer technological data, in particular tool parameters, data on processing modes, workpieces, surfaces, etc., to an external database. With this approach, CAD can become a part of an enterprise PDM-system, which also includes an expert system.

Integration with the CAM system of PowerMILL was realized with the help of the PowerSolution DOTNetOLE.dll. Macros were used to invoke PowerMILL functions, as well as PowerMILL dialog boxes. The structure of an external database for storing technological information has been developed and partially implemented. The application is written in VB.NET.

The control program can be transferred to the CNC machine after the process is formed in the impCAD module.

The working tool-model, the manufacturing technique of which was developed in CAD on the basis of PowerMILL, is made of fluoroplastic on the machine tool with CNC HERMLE C40U. The rest of the parts are made on a lathe. The matrix is made of polyurethane.

The manufactured universal die tooling for manufacturing implants from sheet blanks by plastic deformation with polyurethane was installed on a hydraulic press with a force of 20 MN (Fig. 8). In the container was placed the lower punch, polyurethane matrix, on which the workpiece cut out along the contour was installed, the model and the upper punch were placed on top. Then the upper punch was loaded with a technological force and deformed the workpiece with polyurethane on the model. Since the model was made of fluoroplastic, to reduce the workloads while working out the technology of obtaining the implant model, we used billets of sheet copper M1 0.2 mm thick, and as an elastic matrix - foamed polyurethane. The obtained samples are shown in Fig. 8.

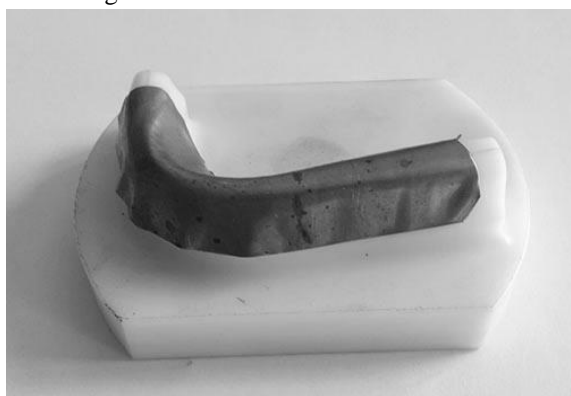


Figure 8. The obtained samples

The experiments showed that the inner surface of the implant fully corresponds to the model, while the outer one requires manual adjustment or correction of the deformation scheme. In general, this technology can be used to produce sheet implants with a complex spatial shape.

V. CONCLUSION

Thus, based on Delcam's PowerSHAPE and PowerMILL products, the technological process of designing and manufacturing implants from sheet blanks was developed and experimentally tested. This process is promising for expanding the nomenclature of implants from sheet blanks with increased mechanical properties and thickness to provide a more rigid structure for connecting damaged bone sites. Preliminary preparation of such implants will significantly reduce the time of the operation. At the same time, design automation makes it possible to more accurately and quickly design and manufacture implants of the required spatial shape. An organized treatment conditions based on the clinical conditions and desires of the patient is very important to achieve predictable results with implantable prosthesis.

Saving the natural teeth and fabricating application. The

dental surgeon must familiarize himself with precision.

ACKNOWLEDGMENT

The author gratefully acknowledge funding from the Ministry of Education and Science of the Republic of Kazakhstan under the target financing program for the 2017-2019 biennium by the program "Production of titanium products for further use in medicine".

REFERENCES

- [1] W. Khan, E. Muntimadugu, M. Jaffe, A. J. Domb, "Implantable Medical Devices," in *Focal controlled drug delivery*, Springer, 2014, 700 p.
- [2] S.M. Petrov, "Bilateral Cochlear Implantation: Indications for Fitting of the Implants," in *J Med Imp Surg* 2: 114, 2017.
- [3] A.R. Jr. Santos, "Bioresorbable polymers for tissue engineering," in *Eberlin D (ed) Tissue engineering. In-Teh*, Olajnica, 2010, pp 235-246
- [4] W. Khan, S. Farah, A.J. Domb, "Drug eluting stents: developments and current status," *J Control Release* 161:703-712, 2012
- [5] Y. Levy, W. Khan, S. Farah, A.J. Domb, "Surface crystallization of rapamycin on stents using a temperature induced process," *Langmuir* 28:6207-6210, 2012.
- [6] R. Schatzer, A. Krenmayr, M. Kals, C. Zierhofer, "Temporal fine structure in cochlear implants: preliminary speech perception results in Cantonese-speaking implant users," in *Acta Otolaryngology* 130: 1031-1039, 2010.
- [7] C. Morera, L. Cavalle, M. Manrique, A. Huarte, R. Angel, "Contralateral hearing aid use in cochlear implanted patients: multicenter study of bimodal benefit," in *Acta Otolaryngol* 132: 1084-1094, 2012.
- [8] S. Nag, R. Banerjee, "Fundamentals of Medical Implant Materials," in: *Narayan R (eds) Materials for medical devices. ASM International*, 2012, pp 6-16.
- [9] B.D. Ratner, A.S. Hoffman, F.J. Schoen, J.E. Lemons, "Biomaterials science: an introduction to materials in medicine," Academic, San Diego, CA, 2004.
- [10] M.R. Jaff, "Advances in the management of patients with vascular disease," *Expert Rev Cardiovasc Ther* 10:151-153, 2012.
- [11] B. Kretz, E. Steinmetz, R. Brenot, O. Bouchot, "First results of clampless distal anastomosis in peripheral vascular bypass with LeGoo, a thermoreversible polymer," in *J Vasc Surg* 55:1821-1825, 2012.
- [12] D.W. Kennedy, "The PROPEL™ steroid-releasing bioabsorbable implant to improve outcomes of sinus surgery," *Expert Rev Respir Med* 6:493-498, 2012.
- [13] G. Mani, M.D. Feldman, D. Patel, C. Agrawal, "Coronary stents: a materials perspective," in *Biomaterials* 28:1689-1710, 2007.
- [14] J. Geller, "CDRH issues nonapproval letter to EES for SEDASYS system," *J Clin Eng* 36:88-91, 2011.
- [15] X. Chevalier, J. Jerosch, P. Goupille, N. van Dijk, F.P. Luyten, D.L. Scott, "Single, intraarticular treatment with 6 ml hylan GF 20 in patients with symptomatic primary osteoarthritis of the knee: a randomised, multicentre, double-blind, placebo controlled trial," *Ann Rheum Dis* 69:113-119, 2010.
- [16] R.D. Laske, D. Veraguth, N. Dillier, A. Binkert, D. Holzmann, "Subjective and objective results after bilateral cochlear implantation in adults," in *Otol Neurotol* 30: 313-318, 2009.
- [17] S.M. Petrov, A.A. Tsjuk, "Instruction for audiologists and cochlear implanted patients," in *J Med Imp Surg*, 2015.
- [18] H. Nagele, M. Azizi, S. Hashagen, M.A. Castel, S. Behrens, "First experience with a new active fixation coronary sinus lead," *Europace* 9:437-441, 2007.
- [19] J.M. Junkins-Hopkins, "Biologic dressings," in *J Am Acad Dermatol* 64:e5-e7, 2011.
- C. E. Misch, "Contemporary Implant Dentistry," Elsevier eBook on Intel Education Study, 3rd Edition, 1120 p., 2008.

Software Refactoring – an Approach Based on Patterns

Violeta Bozhikova*, Mariana Stoeva*, Bozhidar Georgiev*, Dimitrichka Nikolaeva* and Márta Seebauer**

*Department of Software and Internet Technologies, Technical University – Varna, Varna, Bulgaria

**Alba Regia Technical Faculty, Obuda University, Hungary

Abstract—The paper presents an approach for software development based on patterns. On one side, these are patterns in the role of best practices for software development on the other – patterns as bad solutions that must be avoided. Refactoring is a general way to transform a bad solution in a better one. This is a process of source code restructuring with the goal to improve its quality characteristics without changing its external behaviour. In refactoring we replace one software solution with another one that provides greater benefits: code maintainability and extensibility are improved, code complexity is reduced. The developed approach is implemented in software system that can be successfully used both in the real software engineering practice and in software engineering training process.

Key words—Software Refactoring, Software Design Pattern Generation, AntiPattern Identification

I. INTRODUCTION

The paper proposes an approach for software development based on both: Anti-Patterns detection and Design Patterns identification and generation. The presence of Anti-Patterns and Design Patterns is recognized as one of the effective ways to measure the quality of modern software systems. Patterns and AntiPatterns are related [1]. The history of software production shows that Patterns can become AntiPatterns. It depends on the context in which a Pattern is used: when the context become inappropriate or become out of date than the Pattern becomes AntiPattern. For example, procedural programming, which was Pattern at the beginning of software production activity, with advances in software technology gradually turned into AntiPattern. When a software solution becomes AntiPattern, methods are necessary for its evolution into a better one. Refactoring is a general way for software evolution to a better version. This is a process of source code restructuring with the goal to improve its quality characteristics without changing its external behaviour. In refactoring we replace one software solution with another one that provides greater benefits: code maintainability and extensibility are improved, code complexity is reduced ([2]–[10]).

Based on the approach proposed a web system was developed. The system can be used as instrumental tool in the real practice of software production as well in the teaching process - to support several software engineering disciplines in “Software and Internet technologies” Department of the Technical University in Varna. The final effect of its application is to improve the software

quality. It relies on techniques that generate the structure of Software Design Patterns, find AntiPatterns in the code and perform Code Refactoring.

Next section of this paper comments the structure and the basic components of the proposed approach. After, the software implementation of the approach is discussed; the system’s architecture and the basic structural elements are presented.

II. APPROACH FOR SOFTWARE DEVELOPMENT BASED ON PATTERNS AND REFACTORING

The general structure of our approach is presented in figure 1. It takes as input the software source code that has to be refactored. The output is refactored code. The approach relies on accumulation of knowledge about the best practices in programming so “Accumulation of Knowledge” is one of the processes that are performed in parallel with other processes. The refactored code is result of "AntiPatterns Identification and fixing" and "Design Patterns Generation". Before generate Design Patterns it is necessary to analyze the code with the goal to find Design Patterns candidates. The proposed approach comprises the following main component:

A. Accumulation of Knowledge

Aims to provide information on design patterns and AntiPatterns. It contains information about creation, behavioral and structural design patterns and software AntiPatterns in software development and software architecture. Describe the problems that each design pattern solves, the advantages that it provides and the situations in which it is used. For the AntiPatterns - the nature of the problems and possible options for their avoidance are described.

B. Design Pattern Generation

Provides functionality to generate sample implementations of the design patterns structures; basic elements and relationships between them are generated, it’s not implementation of the solution of specific problem. To generate a template user must select appropriate names for key elements. Appropriate names for the key elements must be given by the user, in order to generate a pattern.

C. Refactoring Component:

Provides methods for automatic code refactoring. The code is supplied as input of any method, as for inputs are accepted only properly constructed classes. Each method

performs the appropriate changes and returns the modified code as a result.

D. AntiPatterns Identification and fixing

Provides methods for code analysis. The code is supplied as input of a particular method and as result of code analysis the poorly constructed sections of code are colored. The colored code should be rewritten in order to increase its readability and maintenance.

E. Design Pattern Identification:

Provides methods to examine source code and to identify candidates for design patterns [6]. This component is still under development. Our detection strategy is based on the code inspection. Extensive research has to be conducted to develop techniques to automatically detect candidates of DP in the code.

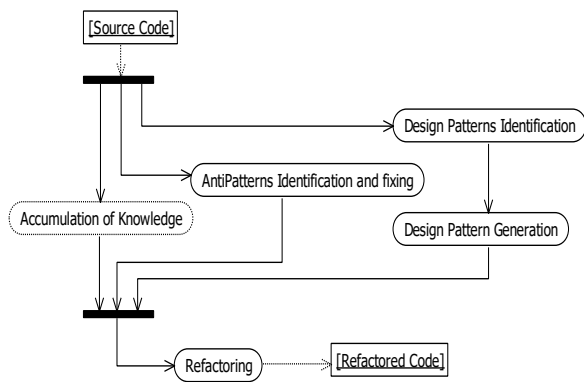


Figure 1. General structure of the approach for software development based on Patterns and Refactoring

III. SOFTWARE IMPLEMENTATION OF THE APPROACH

Based on the approach proposed a web system was developed. The basic structural elements of the web system are presented in figure 2.

A. Main Page:

It aims to present the different sections of the system with a short description and redirect the user to any of them.

B. Encyclopedia:

It aims to provide information on design patterns and AntiPatterns. It consists of two parts: menu type accordion and informative part. The user can select from the menu a concrete DP or AntiPattern. When you choose a concrete pattern then the information about it is displayed in the informative part. The section describes the problems that each design pattern solves, the advantages it provides and the situations in which it is used. For AntiPatterns – their nature and options to be avoided are described. This section is realized as one page with dynamic content that is changed through asynchronous AJAX requests to the server.

C. Design Pattern Generation:

This section offers functionality to generate sample implementations of the structure of the design patterns.

The user chooses a type of pattern, inputs its parameters and click button "Generate". The generated code is displayed below the form. An example of design pattern (Template Method) generation is presented in figure 3.

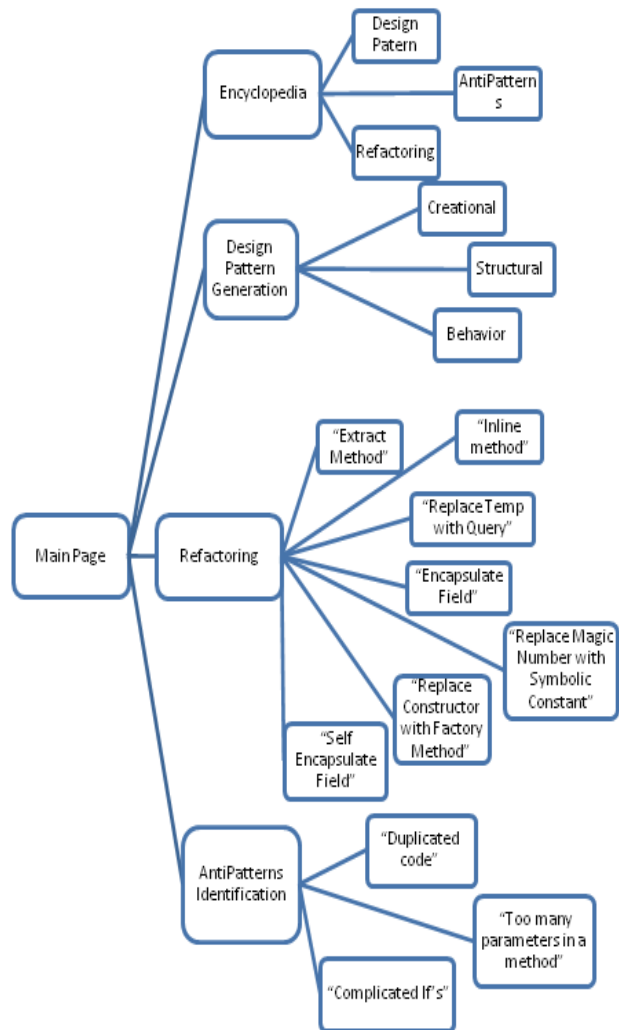


Figure 2. Basic structural elements of the web system

D. Refactoring:

This section provides 8 methods for automatic code refactoring: "Extract Method", "Inline method", "Replace Temp with Query", "Encapsulate Field", "Replace Magic Number with Symbolic Constant", "Replace Constructor with Factory Method" and "Self Encapsulate Field". Each method performs the appropriate changes of the code and the modified code is returned as a result. 8 refactoring methods are provided by the tool. In the left section of the refactoring window (figure 4), the user puts the code, which must undergo refactoring. After entering the necessary parameters the user has to press button "Refactor". The refactoring code is displayed in the right pane.

E. AntiPattern Identification:

This section offers methods for code analysis with the goal to detect AntiPatterns ("Duplicated code", "Too many parameters in a method", "Complicated If's"). The code is supplied as input to each method, which analyzes

and paints the poorly constructed code sections. Then poorly constructed pieces of code must be rewritten to improve code readability and maintenance. Selecting a method (for example “Duplicate code”) a page for entering the code for analysis is visualized. After entering the necessary input data the user must press the button “Identify”. The program will process the code and will paint the problematic code parts in red (the result is shown in the right section – figure 5).

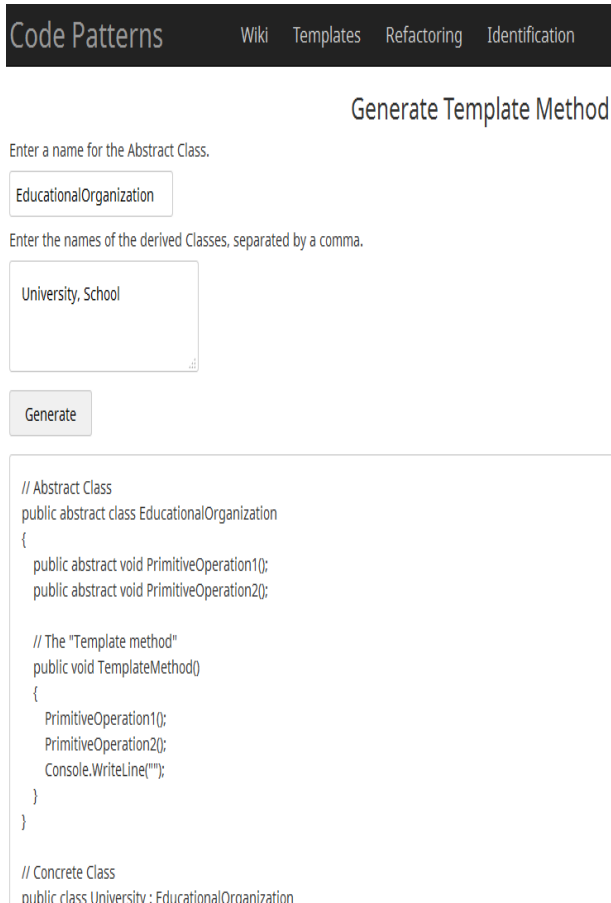


Figure 3. Template Method generation

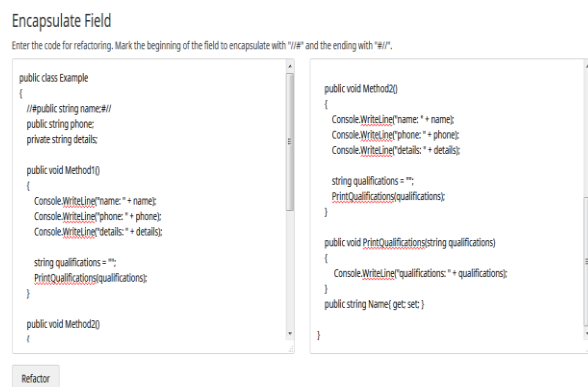


Figure 4. Refactoring window - method “Encapsulate Field”

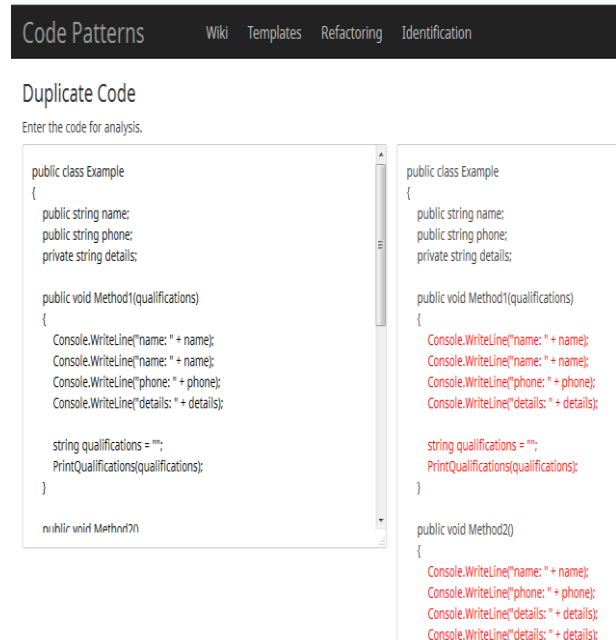


Figure 5. AntiPattern Identification window – “Duplicate code” identification

IV. CONCLUSIONS AND FUTURE WORK

The practical application of the developed software model in the practical exercises on the course “Computer Organization” has led to the following conclusions:

An approach for software development based on AntiPatterns detection and Design Patterns identification and generation is proposed in this paper. It relies on techniques that generate the structure of Software Design Patterns, find AntiPatterns in the code and perform Code Refactoring. Refactoring increases the software quality, it is a general way for software evolution to a better version.

The developed approach is implemented in a software system that that already has been applied in software engineering teaching process but could be also used in the real software engineering practice. It relies on the realization of 4 main sections: educational section that gives information on design patterns and AntiPatterns; Design Pattern generation section that offers functionality to generate the structure of 26 design patterns (Creational DP, Structural DP and Behaviour DP) in C#; AntiPattern identification section that at this time provides realization only of 3 methods for AntiPatterns detection; Refactoring section that provides 8 methods for automatic code refactoring. Each method performs the appropriate changes of the code and the modified code is returned as a result.

Our work associated with the approach presented and the system developed is still in its initial phase. We plan to add new patterns and AntiPatterns in encyclopaedic part. Support for languages other than C# can be provided by the Design Pattern generation Component. Future work is needed to implements more refactoring, AntiPattern and design pattern generation methods. In this time “Design Pattern Identification” section is only sketched and has not been studied and fully developed. So, we plan quite extensive research for the future implementation of this section.

REFERENCES

- [1] William J. Brown, Raphael C. Malveau, Hays W. McCormick III, Thomas J. Mowbray, *AntiPatterns. Refactoring Software, Architectures, and Projects in Crisis*, John Wiley & Sons, Inc., 1998, Canada
- [2] https://en.wikipedia.org/wiki/Systems_development_life_cycle
- [3] https://en.wikipedia.org/wiki/Code_refactoring
- [4] Ali Ouni, Marouane Kessentini, Houari Sahraoui, Mel Ó Cinnéide, Kalyanmoy Deb, Katsuro Inoue, A Multi-Objective Refactoring Approach to Introduce Design Patterns and Fix Anti-Patterns, <http://sel.ist.osaka-u.ac.jp/lab-db/betuzuri/archive/990/990.pdf>
- [5] Martin Drozd, Derrick G Kourie, Bruce W Watson, Andrew Boake, *Refactoring Tools and Complementary Techniques*, <https://pdfs.semanticscholar.org/ae2a/5ccaf697880cb386046e8882a6c268e83312.pdf>
- [6] Jagdish Bansiya, *Automating Design-Pattern Identification*, Dr. Dobb's Journal, 1998, <http://www.drdobbs.com/architecture-and-design/automating-design-pattern-identification/184410578>
- [7] S. Negara, N. Chen, M. Vakilian, R. E. Johnson, and D. Dig, A comparative study of manual and automated refactorings, in *27th European Conference on Object-Oriented Programming (ECOOP)*, 2013, pp. 552–576
- [8] M. Kim, T. Zimmermann, and N. Nagappan, A field study of refactoring challenges and benefits, in *20th International Symposium on the Foundations of Software Engineering (FSE)*, 2012, pp. 50:1–50:11
- [9] Xi Ge and Emerson Murphy-Hill. *Manual Refactoring Changes with Automated Refactoring Validation*. In *Proceedings of the Int. Conf. on Soft. Eng. (ICSE)*, 2014
- [10] Ioana Verebi, *A Model-Based Approach to Software Refactoring*, https://www.researchgate.net/publication/281686403_A_Model-Based_Approach_to_Software_Refactoring

Registration of Changes in the Environment During a Total Solar Eclipse Using ARDUINO

György Hudoba

Óbuda University, Alba Regia Technical Faculty, Institute of Engineering, Székesfehérvár, Hungary
Hudoba.gyorgy@amk.uuni-obuda.hu

Abstract—In August 21, 2017 a total solar eclipse was visible across the United States. I registered the changes in light, temperature, air pressure and relative humidity changes during the eclipse. In the followings I shortly present the device and the results as well.

I. INTRODUCTION

On Aug. 21, 2017, skies darkened from Lincoln Beach, Oregon to Charleston, South Carolina. The event was the first total solar eclipse visible from coast to coast across the United States in 99 years. A total solar eclipse occurs when the shadow cast by the moon is sweeps through the surface of the Earth (Figure 1.). Seeing from Earth, the disk of the moon appears to completely cover the disk of the sun in the sky. During a total solar eclipse, the sun's tenuous outer atmosphere, the corona, becomes visible.

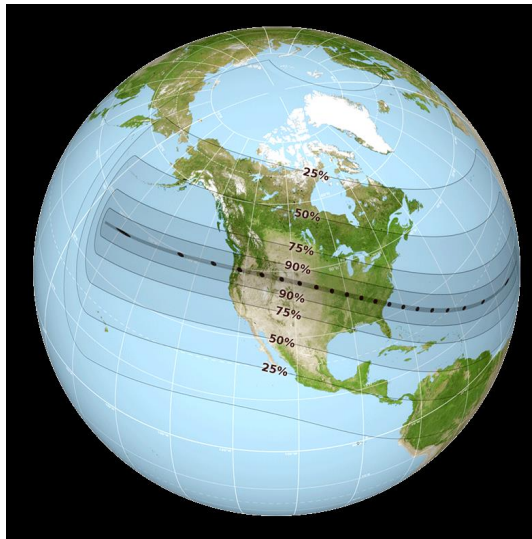


Figure 1. The globe view of the path of totality for the August 21, 2017 total solar eclipse.

[Image Credit: NASA's Scientific Visualization Studio]

My goal was registering and analyzing the changes in the environment conditions, namely the changes of light, temperature, humidity and atmospheric pressure during the eclipse. For this purpose I designed an ARDUINO-based device.

II. THE ARDUINO FAMILY

Arduino is just one small part of the single-board computing (e.g. Raspberry PI among many others) and the

world of embedded electronics and DIY (Do It Yourself) technology. Some selected resources: [1] – [22].

Arduino was named after a bar frequented by students at the Interaction Design Institute in the northern Italian city of Ivrea. The bar was named for an Italian king, Arduin of Ivrea, who briefly ruled Italy around 1000 CE. The word "Arduino" roughly translates to "strong friend."

Arduino is:

- an open-source electronic prototyping platform based on flexible easy to use hardware and software, that was designed for artists, designers, hobbyists, hackers, newbies, or even professionals, and anyone interested in creating interactive objects or environments.
- consists of both a physical programmable circuit board (often referred to as a microcontroller) and a piece of software, or IDE (Integrated Development Environment) that runs on our computer, used to write and upload computer code to the physical board.

There are many varieties of Arduino boards that can be used for different purposes. The term "ARDUINO" refers to a whole family, see TABLE I. below.

The boards differ in sizes and shapes as well as

TABLE I.
ARDUINO MEMBERS

<i>General purpose ARDUINO boards</i>	
	MCU only boards Combined MCU / MPU boards
<i>Special purpose ARDUINO boards</i>	
	ARDUINO Esplora ARDUINO Robot
<i>ARDUINO compatible boards</i>	
	Intel Galileo, Gen 2 Intel Edison
<i>ARDUINO shields</i>	
	Ethernet, WiFi, GSM, Motor, Relay and others

capabilities. Some are credit card sized, others just a tiny stripe, or even round for fashion designers. (The LilyPad line of wearable Arduino boards feature large, sewable tabs for connecting project with conductive thread, and a distinct lack of corners so they don't get caught up in our fabric.)

Most Arduino boards built around an ATmega microcontroller unit (MCU), which is like a complete computer. It has CPU, RAM, Flash memory, A/D converter, and input/output pins, all on a single chip. It is designed to attach all kinds of sensors, LEDs, small motors and speakers, servos, etc. directly to these pins, which can read in or output digital or analog voltages between 0 and 5 volts. Its wide voltage tolerance and low power consumption makes it perfect for the Arduino.

Arduino can interact with buttons, LEDs, motors, speakers, GPS units, cameras, the internet, and so on. This flexibility combined with the fact that the Arduino software is free, the hardware boards are pretty cheap, and both the software and hardware are easy to learn has led to a huge variety of Arduino-based projects.

A. Some special ARDUINO terms

- shield: Arduino shields are modular circuit boards that piggyback onto Arduino to expand with extra functionality.
- sketch: Arduino programs are called sketches. They are usually written in C++.

B. ARDUINO programs

The structure of the sketches is very simple. It consists two parts: *setup* and *loop*.

- *setup*: It is called only once, when the Arduino is powered on or reset. It is used to initialize variables and pin modes.
- *loop*: The loop function runs continuously till the device is powered off. The main logic of the code goes here. Similar to while (1) for micro-controller programming.

The Arduino connects to the computer via USB, where we program it in a simple language (C/C++) inside the free Arduino IDE (Integrated Development Environment) by uploading the compiled code to the board. Once programmed, the Arduino can run with the USB link back to our computer, or stand-alone without it — no keyboard or screen needed, just power.

III. THE DEVICE

For logging the data measured during the total solar eclipse I used the following components:

- ARDUINO Uno R3 board (Figure 2.)
- XD-05 Arduino Data Logging Shield Module (Figure 3.)
- YURobot Easy Module Shield v1 (Figure 4.)
- BMP-280 Barometer Atmosphere Pressure Sensor module (Figure 5.)
- VK2828U7G5LF GPS Module (Figure 6.)

A. The ARDUINO Uno R3 board

The ARDUINO UNO R3 is a General purpose MCU only board. The heart of the board is the Atmel AVR ATmega 328. The ATmega328 on Arduino Uno contains a modified Harvard architecture 8-bit RISC processor as well as a block of flash memory, multiple timers, analog-to-digital converters and PWM generators, all packed into

that one little chip. (The R3 refers to the development phase of the basic ARDUINO board.)

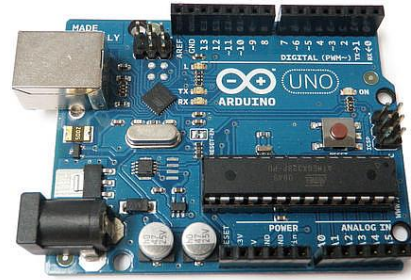


Figure 2. The ARDUINO Uno board used in the project

B. The XD-05 Arduino Data Logging Shield Module

This shield has a RTC (Real Time Clock) module and a SD card module. In addition to, there is an empty space for prototyping. The main features:

- RTC with battery
- Realtime reading
- SD card interface
- Could save data to any FAT16 / FAT32 SD card
- 3.3V level transmit circuit

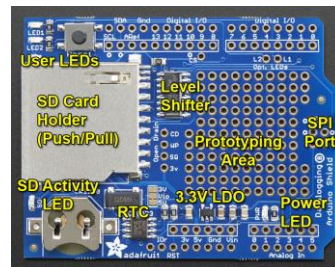


Figure 3. The Data Logging Shield

C. The YURobot Easy Module Shield v1

This shield integrates various module functions and we can directly program to complete the experiment without welding and coping jungle of cables.

Board Features:

- Two pushbuttons
- Two channels LED module (a Blue and a Red LED)
- Full color LED module (one RGB LED)
- Infrared receiver module
- Brightness sensor module (CdS Photorezistor)
- LM35D temperature sensor module
- Passive buzzer module
- Rotary potentiometer module
- DHT11 temperature and humidity sensor module¹
- One I2C interface (SDA A5, SCL A4)
- One TTL serial port
- Two channel digital quantity port (D7, D8)
- One channel analog port (A3)
- Reset button

¹ The DHT-11 Digital Temperature and Humidity Sensor is laboratory pre-calibrated, and uses Single-wire communication



Figure 4. The YURobot Easy Module Shield v1

D. The BMP-280 Barometer Atmosphere Pressure Sensor module

This module has two sensors, one for temperature and another for barometric pressure, and can even be used in both I2C and SPI. (SPI = Serial Peripheral Interface is a synchronous serial data protocol used by microcontrollers for communicating with one or more peripheral devices.) The sensors are very precise: measures barometric pressure with ± 1 hPa absolute accuracy, and temperature with $\pm 1.0^\circ\text{C}$ accuracy. Because pressure changes with altitude, and the pressure measurements are so good, we can also use it as an altimeter with ± 1 meter accuracy.



Figure 5. The BMP-280 module

E. The VK2828U7G5LF GPS Module

I used this GPS module for obtaining the precise geographic coordinates (latitude, longitude and altitude) and the time of the observation.

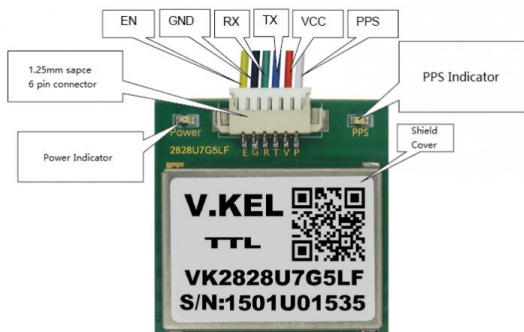


Figure 6. The GPS module

The GPS module communicate with the ARDUINO via serial port.

IV. COMPLETING THE HARDWARE

Although the Data Logging and Easy Shield modules match perfectly to ARDUINO as a piggyback device using pins and headers, completing the device still needed some wiring and other considerations. The BMP-280 module uses I2C, so I connected SDA to the A4, the SCL

to the A5 pin (A means analogue pin of the Microcontroller). The GPS module uses serial communication, so the GPS TX went to the D8, and the Rx to the D7 pin of the ARDUINO board (D means digital pin of the Microcontroller).

For the possible most accurate measurements the A/D converter needs a reference voltage. Because the ARDUINO UNO has limitations, as a circumvention I connected the on board 3.3 V to the pin A3, considered as stabilized value. The idea behind this, that if the power of the microcontroller as reference is changing for any reason, measuring a fixed 3.3 Volt gives us an opportunity correcting the digitized values of other analog devices.

V. OTHER CONSIDERATIONS

As from the previous sections follows, I used many interfaces. Some sensors communicate via I2C, others on one-wire, on serial port or even produced analog signal, and needed free pin for connecting to the A/D converter. Moreover, the A/D converter uses the power level of the microcontroller as a reference.

Powering up the Arduino has two possibilities. One possible way is to use the USB port, the other is using the power jack. Since I want using the ARDUINO as a standalone device, I needed an external power bank. Because the measurement lasted many hours, the input voltage could not be considered as constant. On the other hand, the on board 3.3 V power regulator needs some extra power, so the output power drops about one to two Volts. As a result, if we try to operate the device from 5 Volts, the system become instable, so the external power should be about 9 Volts.

The UNO has some limitations in number of pins and memory. That caused another headache. I could not use the GPS module and the sensors simultaneously. Moreover, as it turned out, the RTC does not keeps the proper time on long run. As a result, the measurement occurred in two steps. First, I had to upload the sketch for the GPS to synchronize the RTC, and get the geographic position of the site, then upload the data collecting and recording sketch.

VI. THE SOFTWARE

The program using the device consists of two sketches: the "GPS2RTC_Sync.ino" and the "EclipseLOG.ino". (The ".ino" extension is the standard for the ARDUINOs.) The first program reads the standard NMEA sentences from the GPS receiver, and prints to the serial monitor the RTC time, the GPS time and date, and the quality of the reception. If the quality of the reception is good enough (Fix = 1, Quality = 2), then the RTC time will be updated (synchronized) once. The serial monitor will show the Fix, and Quality, the geographic position (latitude, longitude and altitude), and the number of GPS satellites seen in each seconds.

The second program at start initializes the SD card, and creates a new .csv file (LOG_xx.CSV), by incrementing the index. After that it creates a record with the following contents:

- # of record
- milliseconds elapsed from the start
- seconds elapsed from 01.01.1970

- the current date and time (UTC) in human readable format
- luminosity (Volts on CdS photoresistor)
- atmospheric pressure (BMP-280)
- temperature in Celsius (BMP-280)
- temperature (Volts on LM35)
- temperature in Celsius (converted from LM35)
- temperature in Celsius from DHT11 sensor
- relative humidity in % from DHT11 sensor
- Ref3 – digitalized value of the stabilized 3.3 V

The registration is started for pushing either K1 or K2 buttons on the Easy Module Shield. (We can stop recording data for the same action.) The data is updated in each second, and is echoed to the serial monitor as well. The records created are stored in a buffer, and after collecting 10 records the content of the buffer will be stored on the SD card.

VII. RESULTS

The measurement was successful. The results (the content of the csv file created) were processed in EXCEL, and present them in graphical form.

On board RTC time:	2017/8/21 15:31:4
GPS time:	15:31:4.0
Date (yyyy.mm.dd):	2017.08.21
Fix:	1
Quality:	2
GPS coordinates:	4351.5336N, 11232.2119W
Coordinates for Google Maps:	43.8589, -112.5369
Elevation:	1463.70 m
# of GPS satellites:	12

Figure 7. The data from GPS module

The Figure 7. shows the data obtained from the GPS module. Figure 8. presents, how the illumination, the temperature and the air pressure changed during the 4 hours observational period, the measurement was taken. The illumination came from the CdS photoresistor, the temperature and the pressure from the BMP-280 detector.

On Figure 9. we can see the temperature obtained from LM35 and the DHT11 sensor, and the relative humidity changes. We have all together three temperature curves. They have the same figure, but different values. This is not an error, but result of different position. The BMP-280 was inside the device, between two shields, the LM35 was exposed to the direct radiation of the Sun, and the third was in a shaded place, under the DHT11 protecting housing.

Zooming into the totality (Figure 10.), we get a nice smooth curve. The bottom of the light curve is not flat, because of the Sun's although dim, but well visible outer atmosphere. the corona. Differentiating the light curve, we can determine the duration of the totality which is 2 minutes 17 seconds, corresponds to the prediction for the site of the observation.

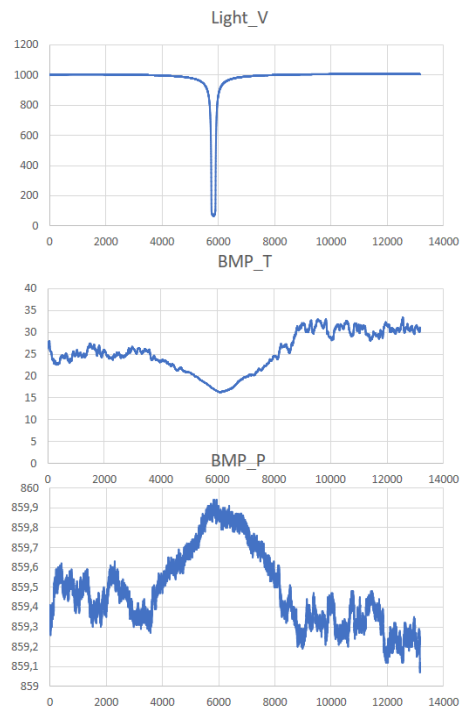


Figure 8. Changes of light, temperature and atmospheric pressure during the 4 hours observational period

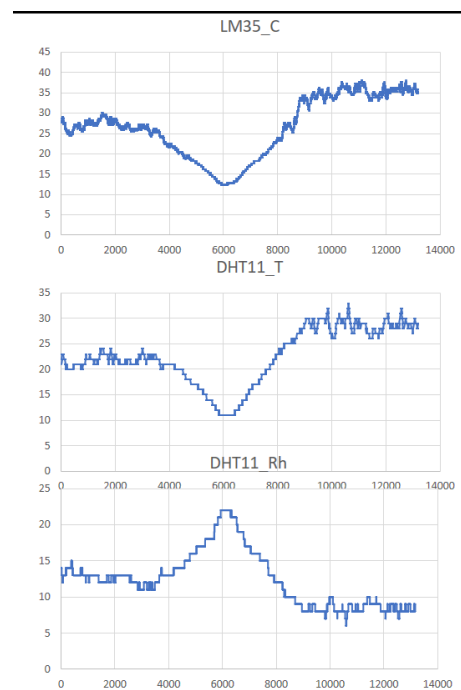


Figure 9. Changes of temperature and relative humidity during the 4 hours observational period

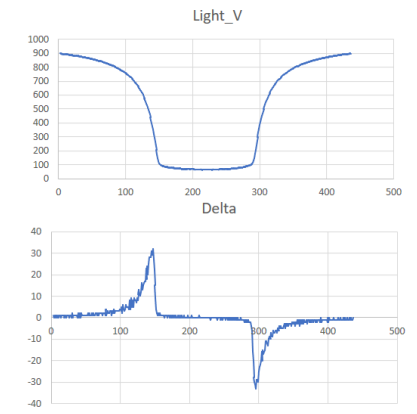


Figure 10. Zooming into the totality. The lower curve is the derivative of the light curve.

RESOURCES

(The internet sources last visited in October, 2017)

- [1] [adafruit-data-logger-shield.pdf](#)
- [2] [Arduino-A000066-datasheet.pdf](#)
- [3] [InstallingAdditionalArduinoLibraries.pdf](#)
- [4] [tavr_arduino_notebook.pdf](#)

- [5] <http://arduinolearning.com/code/arduino-easy-module-shield-v1.php>
- [6] https://cdn.sparkfun.com/assets/home_page_posts/2/0/6/6/arduino_field_guide_copy.pdf
- [7] <https://eclipse2017.nasa.gov/eclipse-who-what-where-when-and-how>
- [8] https://en.wikipedia.org/wiki/Solar_eclipse_of_August_21,_2017
- [9] <http://forefront.io/a/beginners-guide-to-arduino>
- [10] https://github.com/adafruit/Adafruit_Sensor
- [11] https://hci.rwth-aachen.de/tiki-download_wiki_attachment.php?attId=1909
- [12] <https://learn.sparkfun.com/tutorials/what-is-an-arduino>
- [13] <http://physics.kg.ac.rs/fizika/scopes/ARDUINO.pdf>
- [14] <http://students.iitk.ac.in/eclub/assets/lectures/embedded14/arduino.pdf>
- [15] <https://www.arduino.cc/en/Reference/HomePage>
- [16] <https://www.arduino.cc/en/Reference/Wire>
- [17] <https://www.arduino.cc/en/Tutorial/MasterWriter>
- [18] <https://www.arduino.cc/en/Main/Products>
- [19] <http://www.instructables.com/id/Intro-to-Arduino>
- [20] <http://www.instructables.com/id/BMP280-Barometric-Pressure-Sensor>
- [21] <http://www.instructables.com/id/How-to-Use-the-Adafruit-BMP280-Sensor-Arduino-Tuto>
- [22] <https://www.introtoarduino.com/downloads/IntroArduinoBook.pdf>

Bilingual identity transformations in developing bilingualism

Machata Marianna

Alba Regia Technical Faculty, University of Obuda, Székesfehérvár, Hungary

machata.marianna@amk.uni-obuda.hu

Abstract

The paper focuses on the development of a young bilingual's strategic language use and her identity formation in the dual language acquisition process. A functional analysis of her oral language production drawing on naturally-occurring discourse and her metalinguistic comments reveal manifestations of her multiple personality. While analysing the dataset, patterns are identified in her self-regulation process to have a better understanding of how her sense of self is modified and diversified in a dual linguistic environment.

1. Introduction

The present paper is a part of a longitudinal study, which is to investigate and analyze my third child's, Sarah's development in English as a second language (L2) between the ages of one and eleven. At present she is seventeen years old. Sarah has been raised in a dual Hungarian-English context: she has acquired and used these two languages at a time from birth but regarding English she had a limited community support for the simple reason that we are Hungarians and live in Hungary. English is mediated by her parents, primarily by me, her mother, and other native and non-native speakers of English who belong to the family's social network. In order to create favourable conditions for her second language acquisition we try to observe a carefully established language boundary pattern and a clear labour division and language separation between Hungarian and English. It means the use of English and Hungarian is systematically related to certain situations, places and people. My investigation focuses on (1) Sarah's interlanguage development at different levels of language analysis (2) the development of her strategic language use and (3) her identity formation in the dual language acquisition process. I conducted a functional analysis of her oral and some of her written language production (personal letters), drawing on her L1-L2

and L2-L1 language shift and on her narratives retrieved from semistructured retrospective interviews I conducted with her to find evidence of her motives of using L2. To find traces of directly unobservable internal cognitive and psychological processes such as strategies for learning and communication I investigated the communicative intentions in my participant's language choice and metalinguistic comments. While analysing the dataset I tried to identify patterns in her appropriating linguistic forms to functions and also in her self-regulation processes in the dual linguistic environment.

My research is to document (1) how my participant applies code-switching as a strategy to mediate a range of communicative intentions; (2) how her sense of self is reflected in her language alternation and metalinguistic comments; (3) how her language use strategies and her sense of self are modified and diversified in the dual linguistic environment; (4) how her bilingual identity changes across time and situations.

II. The aim of the study

My research aims at understanding my daughter's identity development as it is displayed in language-related episodes of her free and spontaneous interactional exchanges. I present manifestations of her multiple personality and attempt to find important patterns in the data. The excerpts below are to explore how Sarah's interactional practices and orientations to bilingualism are utilized by her as a resource for constituting social relations and identities (Cekaite/ Björk-Willén [2:177]) and how bilingualism becomes a 'significant aspect of self perception and interpretations of developing bilingual life' (Gafaranga [8:510]).

In the present paper the following central research question is addressed:

- 1) How do Sarah's language-related free conversations, code-switches, appeals and orientations to L2 reflect her bilingual identity transformations in her developing bilingualism?

In my category system I rely on the categorization and typology of scholarly research (Baker [1]; Cekaite/Björk-Willén [2]; Cromdal [6]; Gafaranga [8]; Pavlenko [11]). The data presented as excerpts in my dissertation are organized according to the following categories:

1. Social identity development in interaction
2. Self-defence – coping with negative feedback and peer criticism – Asking for justification and reinforcement
3. Highlighting deficiencies and asymmetries in second language knowledge
4. Defining group boundaries – Preserving alliance and privacy
5. Finding ways to enhance learning

III .Research design - Theoretical background

My research falls into the category of qualitative research and involves one person, my own child. It is a single case study conducted longitudinally with a time-span of ten years aiming at understanding a bounded phenomenon: in a marginalized linguistic environment. I focus on the study of language from the point of view of the individual user, putting a special focus on the choices she makes, on the constraints she encounters when using language in social interaction and the effects her use of language has on other participants in the act of communication (Crystal [5]). I approached my data from a purely qualitative perspective, and analysed them with the help of qualitative methods. Interpretive research (Chaudron [3]) is appropriate as I focus on how my participant makes sense of her experiences and also how the researcher in turn makes sense of the data obtained from interviews, observations, narratives, and other sources.

Data for the research were drawn from multiple sources, collected with the help of (1) participant observation and field notes, (2) semi-structured retrospective interviews conducted with the child, and (3) other documents such as the child’s writings e.g. personal letters and drawings. The pragmatic analysis gives opportunity to reveal Sarah’s orientations to bilingualism and allows for knowing more about (1) motives of language alternation and (2) transformations of her identity in her developing bilingualism.

IV. Discussion

I base my discussion on two Gricean assumptions that: (1) communication is a joint activity of the speaker and hearer, which involves the exchange of communicative intentions; (2) a single utterance can convey a range of meanings depending on to whom it is directed and in what context (Grice, [9:50]).

The table below shows the categories as the main organizing principle in grouping the selected discourse pieces of Sarah’s talk.

Table 1. Categories of language-related episodes to reveal Sarah’s multiple identities

Category	Description
1) Social identity development	Affiliation efforts and adjustment to the local norms. Ways of regulating, identifying and redefining herself in the cultural group depending on her personal needs and interests. Orientations to bilingualism at different points of developing bilingualism. Reference to natives’ approval makes a distinction in the local social order.
2) Self-defence – coping with negative feedback - asking for justification and reinforcement	Orientations to peer-initiated criticism, discussions of peer pressure cases. Sarah’s socializing into appropriate ways of regaining entitlement to use a language, which normally does not belong to her monolingual peer group members, her perception of the relative nature of L2 competence, her sense of self in the language learning process.
3) Highlighting deficiencies and asymmetries in second language knowledge	Struggles to reach the respectful position of a sufficiently competent speaker of English. The impact of peers’ comments mean further motivation to use and

	improve her L2. L1 is often used as reference points to assess proficiency in L2 and is a tool to construct knowledge in that language. There is a geographical and social-interactive separation between L1 and L2.
4) Defining group boundaries - Preserving alliance and privacy	Peer group negotiations as social sites for building local social order, values and norms that regulate one another's conduct and group-belonging.
5) Finding ways to enhance learning	Manifestations of Sarah's justification of the importance of L2 knowledge. Her understanding that her English knowledge is an additional asset, which is acknowledged by legitimate, authorized and competent users of L2, such as native peers and school teachers.

1.Social identity development in interaction

The example below reveals Sarah's perception of her L2 competence and that of her position in the local social group.

Excerpt 1

I don't know 'hörghurut' in English, we say only 'ill And even Brandy says so'. (7;2)

The excerpt is suggestive of Sarah's alignment with the L2 language community. The fact that identifying herself with native speakers represented by Brandon authorizes her to be treated as a legitimate L2 speaker (Norton [10]) who is knowledgeable enough to make valid statements

about the target language. Her inferences about her group affiliation betray that in the interaction she had developed a powerful subject position. References to shared language use habits with fully authorized native speakers of L2 like Brendy increases her self-esteem and self-confidence. Building collective identity gives her power and authority.

2.Self-defence – handling negative feedback

Excerpt 2

'Szandra says especially grammar can be learnt only from a teacher, one cannot learn it from one's mother.Then I said that it is possible, I also know it from you, and Kasia is learning from her mother too.' (8;4)

Peers' questioning the relevance of a language learning environment where one's own mother is the mediator of a foreign language and learning occurs in home settings without organized and institutional framework is a recurring topic of Sarah's discourse. The peers' concept of language competence represents a general view of those who base their ideas on institutional learning at school. According to this general view English is identified with the language of schooling, where firm knowledge of words and grammar is the strongest predictor of one's good results in the English lesson. High level of language competence is guaranteed only by institutional learning supported and controlled by an authorized person, preferably a teacher. Sarah's dilemma generated by peer pressure is reflected in her contesting for the position of the competent L2 user (Ricento [12]). She justifies the relevance of learning English at home with her mother arguing that it is as realistic as learning at school with a teacher.

3. Highlighting deficiencies and asymmetries in second language knowledge

Excerpt 3

I didn't know what 'melléknév' in English. I asked Brendy, and he didn't know either, he said if his mother doesn't know something she also looks it up in the dictionary, though she is English. I also said we never say such grammar stuffs, just talk. (9;5)

As academic terms do not constitute the lexicon of the routinely discussed topics in home settings, Sarah has the opportunity to acquire English for only the communicative function but has insufficient knowledge of the language of schooling. Her awareness of the functional differentiation of language use is displayed by her insisting that discrete testing for words is unusual in naturally occurring discourse. Similarly she questions the relevance of word-for-word translation being the primary predictor of one's foreign language proficiency. However, she argues that she has

developed a good level of proficiency regarding the vocabulary of those topics that are involved in naturally-occurring conversations, loose conversations and informal discussions.

4. Defining group boundaries – Preserving alliance and privacy

Excerpt 4

'I don't speak to mummy in the school. Because there we didn't use to. And everybody would stare at me. I don't want to boast.' (9;10)

Sarah's utterance implies that not all of her Hungarian peers belong to the well-informed and initiated circles of her bilingual environment. As a consequence, she judges L2 usage as an insult and disrespect in the L1 environment. She identifies such disrespectful behaviour with boasting, which would entail ousting her from the community. The instance depicts her alignment to the norms of L1 community, where L1 use is the local preference, whereas the use of L2 would be considered as inappropriate.

5. Finding ways to enhance learning

Excerpt 5

Good, but in the test paper it is not enough to write about what I did today. There one must know the material of the lesson! Let's start to learn grammar' (10;7)

The metalinguistic comment in the excerpt above implies that using the target language at home and in other informal frameworks does not allow for success in the language of schooling. There are two distinct functions of language: the communicative and the cognitive function. In free conversations either in formal or informal settings we rely on the communicative function, whereas using the language for academic purposes in particular knowledge areas emphasizes the cognitive function of language. Apparently academic functioning goes beyond naturally occurring free conversational topics and claims for specialized terminology, which requires instructed learning and directed attention.

V. Conclusions

Viewing identity I reported on how Sarah experiences her linguistic and personal identities through the process of her second language acquisition. I attempted to explore what influence bilingualism has on these perceptions, identified and analysed my participant's feelings associated with language alteration. Recurring patterns were studied and evaluated in the silhouettes according to the types of her feelings and self-perceptions.

In terms of her self-perception and identity formulation the applied categories of her bilingual usage shown how she

- (1) consults and involves more competent language users and other authorities of knowledge to determine her linguistic identity and tolerate her imbalances in L2 learning process
- (2) interprets, evaluates and integrates peer pressure and criticism (Cekaite & Björk-Willén [2])
- (3) monitors and evaluates her learning process in terms of L2 in response to her social environment's feedback, and finds opportunities to identify deficiencies and asymmetries in terms of her L2 knowledge
- (4) uses L2 to align with the community to form alliance and privacy or, on the contrary, distances herself from it
- (5) builds distinctions via L2 expertise (Cromdal [6]; Gafaranga [8]) and how L2 expertise reformulates local social order
- (6) seeks and finds opportunities to practise and enhance L2 learning in natural interactions, handles discomfort by finding ways to avert inferior status in terms of L2 (Cekaite & Björk-Willén [2])
- (7) appeals for help and reinforcement to cope with an emerging communication problem

In the investigation of Sarah's communicative intentions and identity transformations I intend to create awareness in my readers that in her language use the indicated categories are not separable. Sarah's language use shows a complex picture where different intentions are interwoven and manifest themselves in a variety of combinations and complement each other. For this reason it is typical that I relate a particular utterance and discourse sample to several communicative intentions and identity perceptions as a consequence, a particular analytical category is exemplified with a number of excerpts. However, if one excerpt is justified to represent more than one type, I followed a twofold principle: (1) I presented that particular excerpt only once, in the most relevant case, or (2) doubledrew on the same excerpt to find underpinnings for two possible categories. Thus certain categories and development stages are exemplified and represented less than others. The reason for such asymmetry is explained by the fact that: (1) I failed to document all the relevant samples during data collection; (2) the representative examples of the analytical categories emerge unevenly and asymmetrically in authentic speech. Such overlaps and disproportions have caused analytical difficulty and my data do not allow for symmetrical and proportionate demonstration of the selected analytical categories. I used unchanged, original data for my analysis to fulfill reliability and validity obligations of qualitative research.

References

- [1] BAKER, C. *Foundations of bilingual education and bilingualism*. 4th ed. Clevedon: Multilingual Matters, 2006.
- [2] CEKAITE, A., & BJÖRK-WILLÉN, P. Peer group interactions in multilingual educational settings: Co-constructing social order and norms for language use. *International Journal of Bilingualism* 17(2) 174-188., 2012
- [3] CHAUDRON, C. Contrasting approaches to classroom research. *Qualitative and quantitative analysis of language use and learning. Second Language Studies*, 19(1), 1-56., 2000
- [4] CRESWELL, J. W. *Research design: Qualitative, quantitative, and mixed method approaches*. Thousand Oaks, London, New Delhi: Sage Publications, Inc., 2003
- [5] CRYSTAL, D. *The Cambridge Encyclopedia of the English Language*, Cambridge University Press, 2003
- [6] CROMDAL, J. Bilingual and second language interactions: Views from Scandinavia. *International Journal of Bilingualism*. 17(2) 121-131., 2013.
- [7] DUFF, P. Research approaches in applied linguistics. In R.B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 13-23). Oxford: Oxford University Press, 2002.
- [8] GAFARANGA, J. Language alternation and conversational repair in bilingual in bilingual conversation. *International journal of bilingualism*. 16(4) 501-527 SAGE, 2012.
- [9] GRICE, H. P. (1957) Meaning. *Philosophical Review*, 67:377-88
- [10] NORTON, B. *Identity and language learning: Gender, ethnicity, and educational change*. Edinburgh Gate: Pearson Education, 2000.
- [11] PAVLENKO, A. *Emotions and bilingualism*. Cambridge: Cambridge University Press, 2006.
- [12] RICENTO, T., Considerations of identity in L2 learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning*, Mahwah, NJ: Lawrence Erlbaum. 2005. pp.895-911.

Working experience and programming language knowledge of university students at Alba Regia Technical Faculty

Zsuzsanna Scherer¹ and Jozsef Halasz^{1,2}

1. Obuda University, AMK, Szekesfehervar, Hungary

2. Vadaskert Child Psychiatry Hospital, Budapest, Hungary

scherer.zsuzsanna@gmail.com

halasz.jozsef@amk.uni-obuda.hu

Abstract— Working experience of university students might have crucial importance in later successful career. The huge market demand for experts in informatics and students in informatics create a special human resource environment. While the players at industry try to find university students in informatics, students are in dire need to establish practical and industrial knowledge.

The aim of the present study was to assess data on working experience in full time university students at Institute of Engineering, Alba Regia Technical Faculty, Obuda University, and to relate it with programming language knowledge.

A self-reported questionnaire was assessed in 173 students from the 193 full time and active students. Data collection was taken place in May, 2016. Sociodemographic data, programming language knowledge, foreign language knowledge, working experience and study progress information were assessed. Statistica 7.0 software package was used to analyze the data, with the help a General Linear Model. For the present analysis, dual training students (obligatory professional job experience from the beginning of their studies) were excluded, thus the data of 155 students were presented.

Earlier or current job was present in the case of 109 students (70.32%), 81 students had current job (52.26%), while 55 students (35.48%) had an actual job related with technical higher education. Programming language knowledge was significantly higher in students with professional job experience ($p < 0.05$). Significant positive correlation occurred between professional technical English language knowledge and programming language knowledge (Spearman $R = 0.31$, $p < 0.0001$). Students with current job report a marked development in their soft skills (communication, presentation skills, decision making, creativity) and their technical knowledge as well.

The interpretation of the above data might be crucial in further establishment of the dual training system.

Keywords: Higher education, Informatics, Programming language, University student, Working experience.

I. INTRODUCTION

A. Working experience of university students

Working experience of university students might have crucial importance in their successful career [1,2]. After graduation, ex-students have to face with the expectations of the job market. Additional to the technical knowledge, working experience and a series of skills are expected at the workplace: flexibility, working alone and in groups, decision making, presentation and communication skills [3-5]. On one hand, regular university studies are not capable providing these kind of experience, and even up-to-date technical knowledge can be found at fast-developing industrial partners instead of the universities. Thus, students are in dire need for having practical industrial air around them. On the other hand, high-tech industrial partners have a continuous problem with the replacement/restructuring of the technical human resources, and graduates in engineering and informatics are among the most successful new job seekers [6-8]. Thus, working as student of a technical university has two driving force, one from the motivational individual surface as successful early career move, and a continuous need at the level of companies for technical newcomers.

As a partial and organized solution to the above question, the dual training system have been introduced to the technical higher education in general, and also to the Hungarian technical higher education system. Dual students have double burden from the very beginning of their studies: additional to the regular university curricula, an official and continuous expectation is also present from the industrial partner [9]. At Obuda University, the dual training was introduced in 2015, at Alba Regia Technical Faculty, in parallel with other Hungarian Universities. The majority of students then did not have official dual training, but a significant number of students had jobs during their university curricula.

To our best knowledge, no official database exist on the working experience of university students in Hungary in general, and at the technical higher education in particular. As a model experiment, the present study addressed the bachelor students of the Institute of Engineering, Alba

Regia Technical Faculty. In the year 2015/16, the faculty had bachelor student in electrical engineering, engineering informatics and technical management. Additional to the working experience, factors related with working and job-seeking behavior were also considered, as motivation, language skills and importantly, programming language knowledge. Higher programming language experience was hypothesized in students with working experience. The development of skills were also hypothesized at university students with working experience.

B. Aims

The aim of the present study was to delineate and analyze working experience and programming language knowledge of university students at the Institute of Engineering, Alba Regia Technical Faculty. The above sample might be used as an important target in our understanding of job preferences and skills of students within technical higher education in general, and informatics related fields in particular.

II. METHODS

The study was approved by Obuda University, Alba Regia Technical Faculty. A detailed questionnaire was assessed in 173 regular students, from the registered and active 193 students (89.64%). The detailed questionnaire was registered by one of the Authors (Z. Scherer), within a personal interview in May 2016. The questionnaire was registered via a person specific code, and the following major data groups were registered: socioeconomic background, social and extracurricular activity at the university, study progress, programming language knowledge, general foreign language knowledge, technical foreign language knowledge, earlier working experience, current working experience, current working experience related with technical higher education. Among the students, 18 students was within the dual training system. As these first year students could be considered as students with present working experience related with technical higher education from the first day of their higher education studies, their results was not included in the present analysis. Altogether, the data of 155 students was presented. Among them, 27 (17.42%) first year students, 58 (37.42%) second year students and 70 (45.16%) third year students participated in the study. The number of students of technical management was 50 (32.26%), the number of students of electrical engineering was 32 (20.64%), while the number of students of engineering informatics was 73 (47.10%). The age of the students was 22.26 ± 1.65 years (mean \pm SD; between 19-27 years).

The self-reported language knowledge in major European languages were assessed, but only the data of subjective English knowledge were presented. The details of the scoring and the extensive description of the questionnaire was described in an earlier manuscript [10].

The following major programming languages were targeted: php, java, html, css, c++, delphi, visual basic (vb), c, c#, r, assembly (ass), sql, pascal, or other programming language. The subjects were binary coded according to their subjective knowledge within a particular programming language.

Statistical analysis. Statistica 7.0 was used to analyze datasets. Additional to the descriptive statistics, a General Linear Model (GLM) analysis was applied, where the year and different type of job experience were used as independent variable, while job experience, general English language knowledge, technical English language knowledge and programming language knowledge were used as dependent variables. Newman-Keuls tests were run in case of post-hoc comparisons. Spearman correlations were also run between programming language knowledge and English language knowledge. The level of significance was set at $p=0.05$.

III. RESULTS

From the 155 students included in the analysis, earlier or current job was present in the case of 109 students (70.32%), 81 students had current job (52.26%), while 55 students (35.48%) had an actual job related with technical higher education. The working experience was highly related with the study progress: students in their later years worked in a higher proportion. In the case of earlier or current work, more than 80% of third year student had a job, while this number was below 40% in the case of students within their first year (Fig. 1, $F_{(1,152)}=11.861$, $p<0.001$). Actual job also showed similar trends, current work was reported over 60% of students within their third year (Fig. 2, $F_{(1,152)}=4.086$, $p<0.02$). Over 40% of students within their third year reported current job in relation with technical higher education (Fig. 3, $F_{(1,152)}=7.976$, $p<0.001$).

Subjective general English language knowledge was not related with the study progress ($F_{(1,152)}=2.111$, $p=NS$), while higher technical English language knowledge was reported in students in their later years of studies ($F_{(1,152)}=3.548$, $p<0.04$).

Among the programming language directions, C#, html and sql was in the first three places (Fig. 4). Programming language knowledge was associated with the study progress (Fig. 5, $F_{(1,152)}=6.934$, $p<0.002$).

The reported programming language knowledge was not associated with earlier or current job ($F_{(1,153)}=1.973$, $p=NS$), nor with current job ($F_{(1,153)}=1.569$, $p=NS$), but was associated with current technical job (Fig. 6, $F_{(1,153)}=6.934$, $p<0.002$).

Significant positive correlation occurred between programming language knowledge and both general English (Spearman $R=0.19$, $p<0.02$) and technical English (Spearman $R=0.31$, $p<0.0001$) language knowledge.

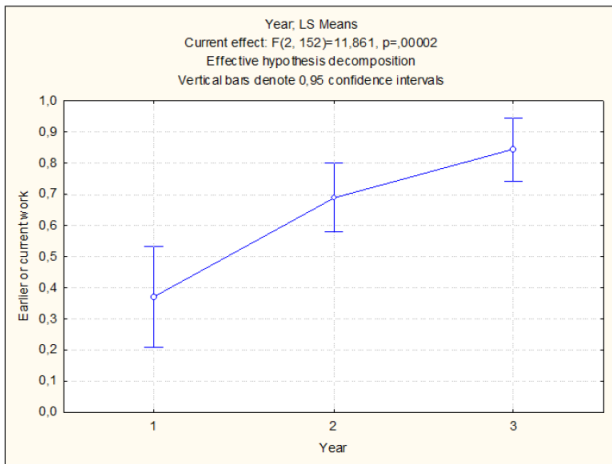


Fig. 1. The description of earlier or current job (0,4=40%) in regular active university students (Institute of Engineering, Alba Regia Technical Faculty) during their study progress. Means and 95% confidence intervals are presented.

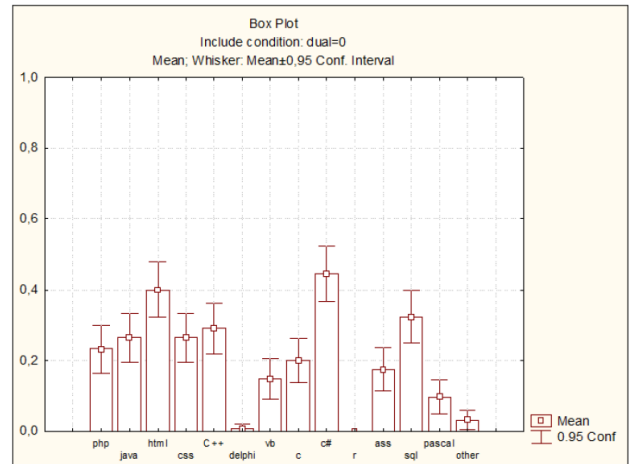


Fig. 4. The description of different programming languages or frameworks (0,4=40%) in regular active university students. Means and 95% confidence intervals are presented.

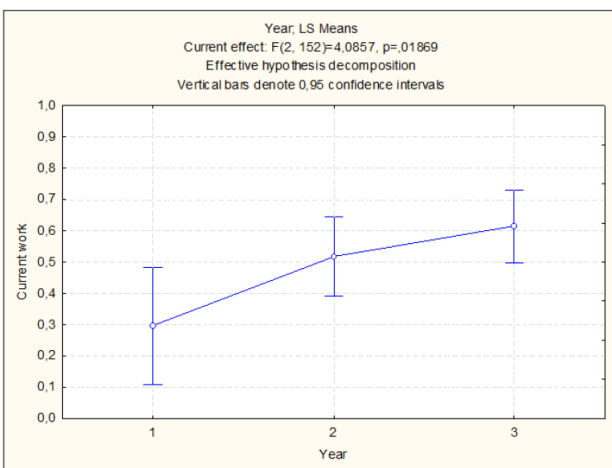


Fig. 2. The description of current job (0,4=40%) in regular active university students (Institute of Engineering, Alba Regia Technical Faculty) during their study progress. Means and 95% confidence intervals are presented.

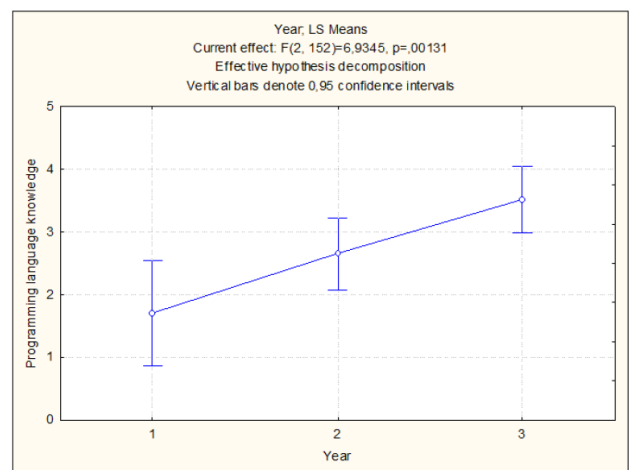


Fig. 5. The number of subjectively reported known and used programming languages. Means and 95% confidence intervals are presented.

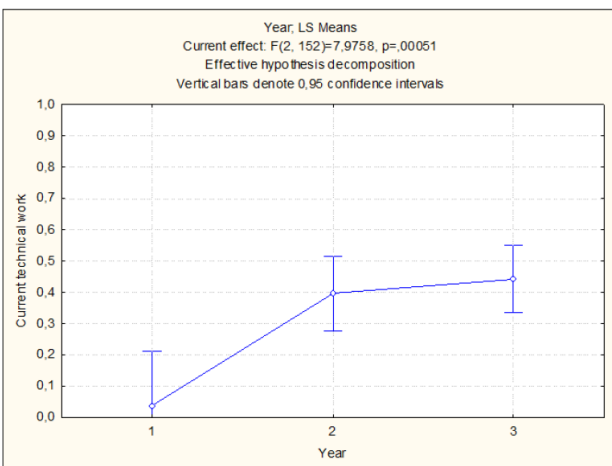


Fig. 3. The description of current technical job (0,4=40%) in regular active university students (Institute of Engineering, Alba Regia Technical Faculty) during their study progress. Means and 95% confidence intervals are presented.

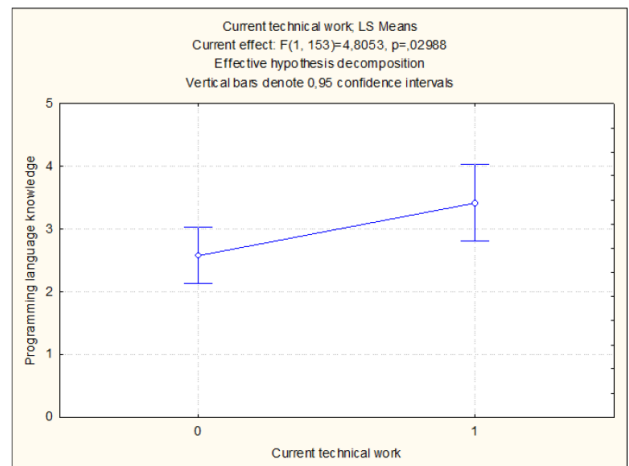


Fig. 6. The number of subjectively reported known and used programming languages. Means and 95% confidence intervals are presented. 0= current technical work is not present; 1= current technical work is present.

The motivation of having current technical job was associated with experience in technical knowledge (41.82%), financial reasons (34.54%), future career (16.36%), and simple possibility (9.10%). During their specific work, the development of problem solving was reported in 72.72%, communication skills in 70.10%, independent working in 43.64%, decision making skills in 41.82%, creativity in 40.00%, flexibility in 36.36%, presentation skills in 27.27%, programming skills in 21.82%.

IV. DISCUSSION

The main results of the present study were the followings. First, job experience was associated with study progress, and the majority of the students had job experience during their studies. Second, programming language knowledge was also associated with the study progress, and current technical job experience was associated with higher programming language knowledge. Third, the students were motivated for technical jobs mainly in order to acquire technical experience, and were reporting significant development in a series of soft skills necessary for their later success.

The proportion of working experience was considerably high in the present sample, and was even higher in the original sample including the students with dual training. Still, the exclusion of the dual students was necessary for the present analysis, as their “success” at the job market was not associated with particular skills acquired at the university. Dual students have to go through a selection process, and specific basal attributes might be different from the beginning, but the authors could not find major differences within the targeted attributes at early level of their university studies [10]. However, differences in the study progress might be present even at their first year of their university studies compared with non-dual students [9].

The present study described three levels of working experience, as previous (or current) work experience, current work experience and current work experience related with technical higher education. General English knowledge was not different in the above groups, but significantly higher technical English knowledge scores were reported in students with current work experience related with technical higher education. Interestingly, similar differences were observed in reported knowledge of different programming languages, and most importantly, a significant positive correlation was observed between reported technical English language knowledge and knowledge of programming languages. This finding was not originally considered as major link between factors necessary for later success in the field of engineering and informatics [4,5].

The development of practical knowledge [3] and the specific support of soft skills [7] are indispensable for later success, and these attributes also occurred in the present study. As a major drive for job experience, the dire need of practical and technical knowledge was the most important

attribute in the motivation for current technical job performance. Students with current technical work experience reported a significant development in their problem solving and communication skills in more than 70% of the cases, and a significant amount reported significant improvement in independent working decision making skills, creativity, flexibility, presentation and programming language skills. Thus, working experience in itself had a major educational power, and its optimal timing and effective intensity should also be reconsidered in students within the dual training system. But, analyzing driving forces might have to include inner motivation, what is a powerful and most important drive in the education of engineering and informatics [11].

The limitations of the study were the followings. Only 155 students were considered in the present analysis, thus further longitudinal and even cross-sectional data collection would be indispensable for high level interpretation. On the other hand, the above sample was representative in the case of Institute of Engineering at Alba Regia Technical Faculty, thus within this context, these data might be interpreted. Other limitation of the study is the self-report of foreign language knowledge and programming language knowledge, additional to the reported development of the soft-skills. These measures can be assessed in a more objective evaluation process, albeit standardized procedures for the interpretation of the above measures in soft skills are still lacking from the literature. In future studies, these issues also should be addressed.

V. SUMMARY

In the present paper, working experience of undergraduate students at Alba Regia Technical Faculty was assessed and interpreted. Current job related with technical higher education was associated with more extensive programming language knowledge. Students with current job report a marked development in their soft skills (communication, presentation skills, decision making, creativity) and the development of technical knowledge as well.

VI. ACKNOWLEDGMENT

The authors declare no conflict of interest. The authors would like to express their gratitude to those university students who participated in the project.

REFERENCES

- [1] M. Beerkens, E. Magi and L. Lill, “University studies as side job: causes and consequences of massive student employment in Estonia”, *Higher Education*, vol. 61, pp. 679–692, 2011.
- [2] P.G. Altbach, L. Reizberg and L.E. Rumbley, “Trends in global higher education: Tracking an academic revolution”, *Report for UNESCO 2009 World Conference on Higher Education*, pp. 1–278, 2009.

- [3] J. Clark and G.W. White, "Experiential learning: A definitive edge in the job market", *American Journal of Business Education.*, vol. 3, pp. 115–118, 2010.
- [4] R. Bridgstock, "The graduate attributes we've overlooked: enhancing graduate employability through career management skills", *Higher Education Research & Development.*, vol. 28, pp. 31–44, 2009.
- [5] T.A. Litzinger, L.R. Lattuca, R.G. Hadgraft and W.C. Newstetter, "Engineering education and the development of expertise", *Journal of Engineering Education*, vol. 100, pp. 123–150, 2011.
- [6] J.M. Laguardo, "Engineering students' academic and on-the-job training performance appraisal analysis", *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 3, pp. 301–305, 2013.
- [7] M.S. Velasco, "Do higher education institutions make a difference in competence development? A model of competence production at university", *Higher Education.*, vol. 68, pp. 503–523, 2014.
- [8] A.M. Cox, M. Al Daoud and S. Rudd, "Information Management graduates' accounts of their employability: A case study from the University of Sheffield", *Education for information*, vol. 30, pp. 41–61, 2013.
- [9] I. Holik and M. Pogatsnik, "An investigation of the effectiveness of dual learning students", *Obuda University E-Bulletin*, vol. 6, pp. 33–39, 2016.
- [10] Z. Scherer, "Description and web-based representation of students' working and employment at Institute of Engineering Alba Regia Technical Faculty", diploma work, Obuda University, 2017.
- [11] J.I. Menges, D.V. Tussing, A. Wihler and A.M. Grant, "When job performance is all relative: How family motivation energizes effort and compensates for intrinsic motivation", *Academy of Management Journal*, vol. 60, pp. 695–719, 2017.

Development of the Robotic Microplasma Spraying Technology for Applying Biocompatible Coatings in the Manufacture of Medical Products

D.L. Alontseva, A.L. Krasavin, A.T. Kadyroldina, A.T. Kussaiyn-Murat, D. M. Nurekenov,
Ye.T. Zhanuzakov, N.V. Prokhorenkova

D. Serikbayev East Kazakhstan State Technical University/ Department of Instrument Engineering and Technology
Process Automation , Ust-Kamenogorsk, Kazakhstan

dalontseva@mail.ru
alexanderkrasavin@mail.ru
akadyroldina@gmail.com
princes__ka@mail.ru
dnurekenov@gmail.com
zhan_erzhan@mail.ru
nadin_kaz@mail.ru

Abstract - The paper describes the algorithms and software development for a robotic technology of microplasma spraying of powder and wire materials for applying biocompatible coatings for medical implants and instruments. The authors observe the challenges and prospects of the development and implementation of the robotic technology for manufacturing medical products.

I. INTRODUCTION

The multi-purpose methods of Thermal Coating Spraying have become popular all over the world lately [1, 2]. One of the major methods of gas-thermal deposition of coatings is plasma spraying. The micro plasma spraying (MPS) method is characterized by a small diameter of a spraying spot (1 ... 8 mm) and low (up to 2 kW) power of plasma, which results in low flow of heat into the substrate [3-5]. These characteristics are very attractive for the deposition of coatings with high accuracy, in particular for applying biocompatible coatings in the manufacture of medical implants.

However, the treatment of surfaces of complex configuration presents a challenge for the implementation of the thermal spraying technology and requires automated manipulations of the plasma source and/or the substrate along with robotic control for appropriate treatment of a surface [1, 2].

At present, robot manipulators are widely used in metallurgical industry, automotive industry and mechanical engineering, allowing to automate the plasma processing. However, they are used only for large-scale production, because every transition to a new product requires complex calibration procedures to achieve compliance with the model set in the robot previously. Thus, the problem of automatic code generation of a robot program for the model specified by means of CAD is in the limelight of researchers and developers of robotic systems [6-8].

The main prerequisites for the development of the research were the analysis of technical issues arising from the industrial robot application for coating by plasma jets, and the desire to expand the scope of tasks solved by the application of an industrial robot. The authors of this paper have carried out a work in the field of application of automated plasma methods of biocompatible or protective coating deposition, described in papers [9-11] and protected by certificates of intellectual property [12, 13]. There is a need to develop methods of plasma coating to create medical products. There is some successful research in the technology development of biocompatible coatings microplasma spraying of biomedical application onto different types of implants conducted by scientists of E. O. Paton Institute of Electric Welding, National Academy of Science of Ukraine [4, 5], which is in close relationship with the present research, because we have used the technological equipment for microplasma spraying developed at E. O. Paton Institute of Electric Welding (IEW), namely the MP-004 microplasmotron mounted on the arm of Kawasaki industrial robot. Currently, on the basis of D. Serikbayev East Kazakhstan State Technical University (EKSTU) there is a robot-manned floor for microplasma materials processing, it allows testing new technological solutions to use biocompatible coatings for medical purposes with high (precision) accuracy. In order to obtain coatings with the desired structure and properties, it is very important to provide accurate modes of deposition and modification of coatings by plasma.

As noted by several researchers [1, 2], the main disadvantages of the coatings achieved by using gas-thermal methods are their high porosity and occasional poor adhesion to the substrate. Porosity can be useful at times, as in the case of ensuring reliable fixation of orthopedic implants into bones on account of the intergrowth into the pores of the bone tissue, etc., but in this case it needs to be controlled.

The aim of this work was to develop a robotic micro plasma spraying technology for applying biocompatible coatings in the manufacture of medical products.

II. EXPERIMENT

A. Equipment and Materials

Within the activities of modern technologies development by D. Serikbayev EKSTU an experimental laboratory industrial complex for plasma treatment of materials based on an industrial robot has been established. Kawasaki RS-010LA (Kawasaki Robotics, Japan) industrial robot is a device consisting of moving parts with six degrees of freedom to move according to a predetermined track. It is controlled by a E40F-A001 programmable controller. MP-004 microplasmotron for applying the powder or wire coating produced by E. O. Paton IEW, Ukraine is mounted on the robot arm. The assembly of the system has been carried out by Innotech LLP, Kazakhstan.

Kawasaki RS-010LA robot manipulator characteristics:

- Number of degrees of freedom - 6;
- Positioning accuracy – 0.06 mm;
- Maximal linear speed - 13100 mm/s;
- Engagement zone - 1925 mm;
- Working load capacity - 10 kg.

The study has dealt with the starting materials for coating deposition: powders, wires and resulting coatings obtained by means of microplasma spraying, as well as substrates (in most cases 3 steel substrates treated by sandblasting were used). The range of materials in the study was broad enough to ensure the mastering of technological processes for different materials. Co-based and hydroxyapatite powders as well as Titanium wires have been used as the main materials for working out the microplasma spraying processes of biocompatible materials.

B. Methods

The technologies of plasma spraying of coatings require accurate adhering to the number of technological parameters (the distance from the plasma system nozzle to the surface of a workpiece, the nozzle movement speed, etc.) during the entire processing time. Exceeding these parameters beyond the permissible limits can lead not only to rejected products, but also to an accident (a short circuit). In cases when the robot program is generated according to a given geometrical model of a processed workpiece or part, the deflection of the shape of the real object from the model often leads to the violation of technological parameters of processing with all its undesirable consequences. This problem is particularly acute in the case of large-sized objects, including medical implants or instruments, when small relative errors of geometric parameters and object positioning correspond to unallowably high absolute deviations of the distances between the tool mounted on the manipulator and the object surface. The radical method of solving these problems is pre-scanning the surface of an object

A modern robot manipulator can be considered as a means of allowing setting spatial position and orientation

of an arbitrary tool with high precision and accuracy. If a distance sensor or a vision system element (camera or projector) is used as a tool, the robot manipulator can be an excellent basis for establishing a system of surface scanning.

The basic idea of the proposed method - the development of a combined system for scanning with the split of scan process into two phases: a rough scan phase and a refining phase. For rough scanning a vision system, that uses a single camera mounted on the manipulator and a fixed structured light projector is supposed to be used. During the rough scan phase, photography of an “illuminated” object from several points of space is performed (with the known orientation of the principal optical axis of the camera). By the images obtained in the shooting process, the software of the scanning system produces a segmentation of the object surface and builds an approximate 3D model of the object. According to the segmentation results, a set of reference points is selected on the surface; and if we know their spatial coordinates, we will be able to construct a 3D model of the object. After selecting reference points, the software generates the program of the manipulator which successively passes the reference points performing surface scanning at each point.

The vision system will be built on the original algorithm, implemented in three stages processing the image obtained by the camera: 1) building a function module of the intensity gradient; 2) constructing a set of lines of this function level (the structuring of the system of level lines radically simplifies the task of finding correlation between the lines obtained in the processing of the two photos taken at different camera positions); 3) calculating spatial coordinates of the scene points, whose images lie on these lines.

The implementation of the proposed algorithm, in contrast to the common computer vision algorithms, does not require much computational power. Besides, the algorithm of level lines is easily parallelized and makes it possible to set up the software processing system on a personal computer (when using the CUDA system for efficient implementation of parallel algorithms).

Thus, we are developing an intelligent automated system of controlling an industrial robot manipulator, which allows a robot arm to move along a given 3D trajectory, a model of the product that the robot will be treating with plasma. A distinctive feature of the proposed system is pre-3D scanning the surface of the rough workpiece or the workpiece in process. We are planning to implement automatic generation of a robot-manipulator program code, taking into account the data of the 3D scanning of an object to be processed, previously held by means of distance sensors mounted on the robot manipulator. This will allow to use workpieces varying in a wide range of geometric parameters and processing products, whose geometrical parameters are determined with low accuracy or products with deviations from a predetermined shape.

III. RESULTS AND DISCUSSIONS

At D. Serikbayev EKSTU the prototypes of coatings from Titanium - wires and Co-based and hydroxyapatite powders have been produced using the robotic complex for microplasma deposition. The parameters for

additional processing of coatings by a plasma jet were selected on the basis of mathematical modeling of the temperature fields arising in the “coating-substrate” system when heated by a travelling plasma source [12]. The experience of getting coating from powders and wire of a range of alloys with the use of this complex was successful; the results were published in [9-11].

To solve the problem of providing the desired trajectory of the plasma source, we have developed the software which converts the drawings made in AutoCAD and Compass to the robot controller by selecting the graphics primitives (line, arc, etc.) from the drawings and transferring them into the commands for the robot arm movement [13]. However, we suppose that this method is unsatisfactory for plasma treatment of the large-sized implants and medical instruments surface, because a 3D model of the surface to be processed is needed to move the robot arm with plasma source accurately during the plasma processing. Currently we are developing an intelligent automated system of controlling an industrial robot manipulator, which allows carrying out product surface hardening treatment: coating application using a microplasma method, and plasma irradiation modification of surfaces of complex shape products. Preliminary 3D scanning of the surface is being processed and generation of the program code is carried out by the same robot manipulator.

As it seems to us, two main categories can be distinguished in the methods of machine binocular vision developed to date: methods based on the detection of image features [14, 15] and methods based on minimizing the energy function [16-18]. Most of the methods of the first group are associated with the so-called problem of edge detection. Edge detection includes a variety of mathematical methods that aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed edges. It should be noted that the methods of the first group find the correspondence between some points of two images. For the methods of the first group it is not necessary to find a correspondence between all the pixels of two images, as well as to find the disparity function that minimizes the so-called energy function.

The method proposed by us is based on finding the correspondence between the curves (contour level curves of modulo of gradient of intensity function) and has some common features with the methods of both the first and second groups.

A black and white image can be considered as a discretization of a continuous function of two arguments $I(x,y)$ (*Intensity function*). In image processing, so-called gradient image processing technique is widely used, based on the numerical calculation of the intensity function gradient. Typically, the magnitude of the gradient vector $F(x,y)$ (1) and its direction are calculated separately as a discrete convolution of the image matrix with one of the specially designed convolution kernels. In our experiments we use the so-called Sobel kernel.

$$F(x,y) = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2} \quad (1)$$

The proposed algorithm is based on finding the correspondence between the level contour curves of two functions of intensity gradient modulo F_1 and F_2 (for the left and right images correspondingly). On the set of level curves of the function, we can introduce a partial order relation, which we denote by the symbol \preceq . We assume that $s_1 \preceq s_2$ if the curve s_1 lies inside the region bounded by the curve s_2 . As a consequence, set of level lines can be represented by a tree, as shown schematically in Fig 1.

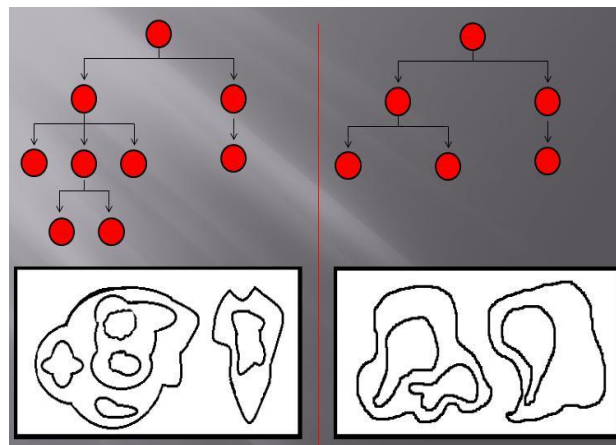


Figure 1. Level Tree

One of the leading ideas of the proposed method is the use of a hierarchical structure of level lines (given by a level tree) to simplify the task of finding a correspondence between the level lines of the functions F_1 and F_2 . In fact, after constructing the level tree, the problem of finding the correspondences between isolines of functions F_1 and F_2 reduces to the problem of identifying isolines that are descendants of a single tree node. Obviously, to implement the methods of this kind, an effective algorithm for constructing isolines is needed. Such algorithms are used both in computer graphics and in image processing, and these algorithms still continue to be improved [19-20].

We have developed an algorithm for constructing the contour curves of the function of two variables that improve modification of the marching triangles classical algorithm. The original Marching Triangle algorithm does not specify any boundary edges processing sequence. It defines only a single pass into the edge list to process all boundary edges with the procedure steps including the estimation of a new potential triangle and its sphere test mesh. According to the implemented data structures and the method to add new edges in the edge list, the edge processing sequence can be different from each other. The resulting mesh from the Marching Triangle depends on the edge processing sequence and it can be different if the sequence is changed. The algorithm developed by us is free from these restrictions and allows parallelization.

Currently, we are developing an algorithm that allows us to identify many isolines. We have developed an effective algorithm for finding a numerical measure of the geometric proximity of two domains bounded by a polygon. The developed method is based on minimization of the energy function, for calculation of which the above-mentioned measure of geometrical proximity of spatial regions is used.

IV. CONCLUSION

Coatings from biocompatible materials deposited by the microplasma according to recommended modes onto steel substrates have been obtained. It is shown that the microplasma spraying method allows applying a wide range of materials: hydroxyapatite, Co-based powders, Titanium – wires. Successful deposition of biocompatible coatings with sustained characteristics on parts of complex shape, which are endoprostheses, requires steady travelling of the plasma source along the sprayed surface of the product. For this purpose, it becomes necessary to equip the deposition unit with a robot manipulator and to develop an intelligent automated system of controlling an industrial robot manipulator, which allows a robot arm to move along a given 3D trajectory.

Multi-view 3D – Reconstruction algorithm has been developed to scan an object quickly. The algorithm is based on finding the correspondence between the isolines of the images intensity gradient functions.

ACKNOWLEDGMENT

The study has been conducted with the financial support of the Science Committee of RK MES in the framework of the program target financing for the 2017-2019 biennium by the program 0006/PTF-17 “Production of titanium products for further use in medicine”.

REFERENCES

- [1] R. C. Tucker, Ed Introduction to Coating Design and Processing, *ASM Handbook, Thermal Spray Technology* Volume 5A, 2013, pp.76–88.
- [2] A. Vardelle, Ch. Moreau, J. Nickolas, A. Themelis, “Perspective on Plasma Spray Technology”, *Plasma Process*, 35, 2015, pp. 491–509. DOI 10.1007/s11090-014-9600-y.
- [3] E. Lugscheider, K. Bobzin, L. Zhao, J. Zwick, “Special Issue: Thick Coatings for Thermal, Environmental and Wear Protection”, *Advanced Engineering Materials*, Volume 8, Issue 7, 2006, pp. 635–639, DOI: 10.1002/adem.200600054.
- [4] Yu. Borisov, I. Sviridova, E. Lugscheider, A. Fisher, “Investigation of the Microplasma Spraying Processes”, *The International Thermal Spray Conference*, Essen, Germany, 2002, pp.335–338.
- [5] A. V. Andreev, I. Y. Litovchenko, A. D. Korotaev, D. P. Borisov, “Thermal Stability of Ti-C-Ni-Cr and Ti-C-Ni-Cr-Al-Si Nanocomposite Coatings”, *12th International Conference on Gas Discharge Plasmas and Their Applications*. IOP Publishing Journal of Physics: Conference Series 652, 2015. DOI:10.1088/1742-6596/652/1/012057.
- [6] E. I. Nelayeva, Yu. N. Chelnokov, “Solution to the Problems of Direct and Inverse Kinematics of the Robots-Manipulators Using Dual Matrices and Biquaternions on the Example of Stanford Robot Arm”, *Mechatronics, Automation, Control*, Volume 16, No. 7, 2015, pp. 456–463.
- [7] M. Rodrigues, M. Kormann, C. Schuhler, P. Tomek, “Robot Trajectory Planning using OLP and Structured Light 3D Machine Vision”, *9th International Symposium*, Greece, ISVC 2013, Part II, LNCS 8034, 2013, pp. 244–253.
- [8] F. J. Brosed, J. Santolaria, J. J. Aguilar, D. Guillomia, “Laser triangulation sensor and six axes anthropomorphic robot manipulator modelling for the measurement of complex geometry products”, *Robotics and Computer-Integrated Manufacturing*, Vol.28, 2012, pp. 660–671.
- [9] D. Alontseva, A. Krasavin, N. Prokhorenkova, T. Kolesnikova, “Plasma – Assisted Automated Precision Deposition of Powder Coating Multifunctional Systems”, *Acta Physica Polonica A*, in press.
- [10] D. L. Alontseva, A. V. Russakova, A. L. Krasavin, N. F. Denisova, N. V. Prokhorenkova “Automated precision deposition of powder multifunctional coatings and microplasma surface treatment”, *Fundamental’nye problemy sovremennogo materialovedeniya (Basic Problems of Material Science)*, Vol. 14, No 1, 2017, pp.88–94. <http://www.nsmnds.ru>
- [11] D. L. Alontseva, A. L. Krasavin, O. B. Ospanov, “Software Development for a New Technology of Precision Application of Powder Coating Multifunctional Systems”, *11th International Symposium on Applied Informatics and Related Areas*, Hungary, November, 2016, pp. 140–143. <http://ais.amk.uni-obuda.hu/>
- [12] A. L. Krasavin, D. L. Alontseva, N. F. Denisova “Calculation of temperature profiles in the two-layer absorbers with constant physical characteristics heated by a moving source”, Certificate of authorship No.0010558 of the Republic of Kazakhstan for the computer program. No.1151 of August 20, 2013.
- [13] D. M. Nurenkov, A. L. Krasavin, D. L. Alontseva, “Converter for DXF drawings into AS language of robot manipulator Kawasaki RS010L”, Certificate of authorship No. 009030 of the Republic of Kazakhstan for the computer program, no 1490 of June 21, 2017.
- [14] C. Schmid, R. Mohr, C. Bauckhage, “Evaluation of interest point detectors”, *Journal of Computer Vision*, Vol. 37(4), 2000, pp. 151–172.
- [15] E. Rosten, R. Porter, T. Drummond, “Faster and better: a machine learning approach to corner detection”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 32, 2010, pp. 105–119.
- [16] T. Meltzer, C. Yanover, Y. Weiss, “Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation” *Tenth IEEE International Conference*, Computer Vision, ICCV, 2005.
- [17] Y. Boykov, V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision”, *IEEE Trans Pattern Anal Mach Intell*, no. 26, 2004, pp.1124–1137.
- [18] V. Kolmogorov, R. Zabih, “Computing visual correspondence with occlusions using graph cuts”, in *Proceedings of the 8th International Conference on Computer Vision*, ICCV, vol. 2, 2001, pp. 508–515.
- [19] T. Lewiner, H. Lopes, A. W. Vieira, G. Tavares, “Efficient Implementation of Marching Cubes Cases with Topological Guarantees”, *Journal of Graphics Tools*, Vol. 8, No. 2, 2003, pp. 1–15.
- [20] Marc Fournier, “Surface Reconstruction: An Improved Marching Triangle Algorithm for Scalar and Vector Implicit Field Representations”, *IEEE XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

DSL for Automatic Control System Modeling of Technological Process

O. B. Ospanov, A.L Krasavin, D.L. Alontseva, A.V. Russakova

D. Serikbayev East Kazakhstan State Technical University/Department of Instrument Making and Automation of Technological Processes, Ust-Kamenogorsk, Kazakhstan

ospanovzhas@gmail.com

alexanderkrasavin@mail.ru

dalontseva@mail.ru

arussakova@gmail.com

Abstract - This paper concerns the description of experimental system of the processes of automated control modeling. The system supports uniquely created DSL and allows describing signals and typed blocks of automation system.

I. INTRODUCTION

Currently acute problem of increasing prices and the reduction of natural resources of non-renewable energy sources, together with negative impact of combustion products emission led to the fact that many research works aim at solving problems of optimization and increase of efficiency of energy usage [1-9].

Primarily it should be noted that most coal burning boilers do not have systems of automatic control of air-fuel mixture concentration. As a rule, feeders, regulating coal dust intake, are chosen empirically and are not regulated automatically, while airline unit has only one automatic control loop to maintain constant air quantity flow (when the flow amount is chosen empirically as well) [1-3]. Nozzles adjustment is usually performed manually, and in fact, empirical evidence had shown that worker's experience and skills could significantly influence the efficiency coefficient of a steam system. At the same time, we are perfectly aware that the efficiency of air-fuel mixture combustion with fixed features of applied fuel dust (coal grade, composition and moisture content in coal) highly depends on air-fuel mixture concentration. Besides, the properties of air fuel are changing during a time of a continuous operation of a boiler, thus, on-going adjustments of fuel mixture making system are ineffective to maintain stable efficiency of a boiler [4,5].

The analysis of attempts to apply the methods of automatic control of the fuel mixture making in boiler and in metallurgical industries, leads to the conclusion that these attempts have been made along two main directions: in some cases, the controlled parameter was the concentration of dust. In other cases – it was the concentration of carbon monoxide or oxygen in the combustion products. Accordingly, in both cases the process of automatic regulation provided with constancy in time of a measured parameter, but not power output. Moreover, engineers for particular operating conditions chose optimal (in average) parameter as a rule of thumb. Obviously, as external factors (chemical and fractional

composition of coal dust, moisture content in coal, etc.) tend to vary, given methods cannot maintain maximum achievable efficiency of a boiler unit during some period of time. Besides, the stabilized parameter was chosen heuristically, as continuous operation mode of boilers is not suitable for large-scale and long lasting experiments. It should be noted that in recent studies of control systems for boilers, based on tracking of several system parameters (multifactorial system) [6-8], but the measured values also do not have a clear and sustainable link with the generated power. Therefore, previously used attempts to set up automated systems of air-fuel mixture making did not make it possible to maintain modes, similar to optimal for a prolonged period. However, even taking into account this fact, their application displayed significant efficiency increase of boilers, equipped with the systems in comparison with boilers, where parameters of air-fuel mixture were regulated manually [2,4,8]. The latter circumstance allows assuming that the proposed methods of active search for optimal of the fuel-air mixture parameters at a fixed heat output can significantly increase the average efficiency of steam systems.

The difference between our ideas from existing analogues:

-firstly, we propose to minimize fuel consumption, using power output as the main measurable parameter, while implemented to date systems of automatic control of the fuel-air mixture making only stabilized parameters in a certain way correlated with the generated power, but in no way did not determine it;

-secondly, we propose a system, that is able to actively search most preferable parameters for air-fuel mixture at a given (fixed) power output unlike previously proposed systems implemented on the basis of classical feedback theory [8];

-thirdly, to decrease system inertia and to enhance its reliability, in contrast to most previously proposed systems where one measurable parameter was used, we suggest to use several measurable system parameters. Along with indirectly measured power output, we suggest to take into account such factors as carbon dioxide in products of combustion and fuel dust concentration in a mixture being made.

The problem of complex automated control systems modeling is relevant due to technical equipment evolution

and increasing complexity of controlled objects. Software modeling as a rule requires high proficiency in the field of software development, that is not always possessed by Engineers usually are not skilled in the sphere of software, though on the other hand, IT specialists lack essential knowledge in the field of system automation and control.

Nowadays the world's practice recognizes two main approaches to this problem solving:

1) Development of visual systems programming, which allows the specialist to create the model of a system using graphical editor by means of combining the blocks – primitives. Simulators of electric schemes (MicroCap, Proteus etc.) can serve as examples of this approach.

2) Modeling system development supports problem-oriented language development.

The main examples of such an approach are the systems of synthesis and verification of logic circuits, that support software programming languages: HDL (Hardware Design Language), Verilog and VHDL (VHSIC - Very high speed integrated circuits Hardware Description Language). In recent years, attempts have been made to create problem-oriented languages for modeling the processes of some specialized areas of industrial automation [9], as well as DSL (Domain Specific Language) for model – driven development of robotics [10].

This work concerns the description of experimental system of modeling of automated control processes. The system supports uniquely created DSL and allows describing signals and typed blocks of automation system. As well as HDL, the suggested DSL allows to create new models by means of primitives' aggregation with description of blocks-chain program as well as by means of module behavior description.

II. METHODS

Simulation is an extremely important tool for all and every control engineer who develops the system of automation and control of technological processes in industry.

For non-linear plants, there is often no alternative for control engineer but only trial-and-error approach, using computer simulation. Thus, hardly any control engineer would not use simulation at least occasionally. Market offers many highly effective special-purpose simulation software tools, e.g. for the simulation of electronic circuits, or for the simulation of dynamics of multicore systems, and there is (or at least used to be) a good reason for that. However, there is no market to offer special-purpose control system simulators, in spite of the fact that control is such an important application of simulation.

The modeling of control systems can be considered as a particular case of modeling dynamic systems. Indeed, in most cases, the model should represent the interaction between environment (plant to be controlled) and the control system. Plants are mostly represented by continuous time systems whose behavior is often described by partial or ordinary differential equations. Therefore, the modeling system should be a continuous time modeling system. On the other hand, digital controllers can be represented mathematically as discrete event systems. These circumstances demand methods

that can deal with heterogeneous components that exhibit a variety of different behaviors.

In many practical cases, the most natural and adequate mathematical model of a plant is a system of differential equations. However, with respect to modeling the control system, we come to a peculiar paradox: mathematically, in such a model, the controlling action is a deterministic function of time, whereas in reality we cannot predict control function of action before the modeling process being carried out. For example, let's consider classical problem of control of an inverted pendulum (Fig. 1).

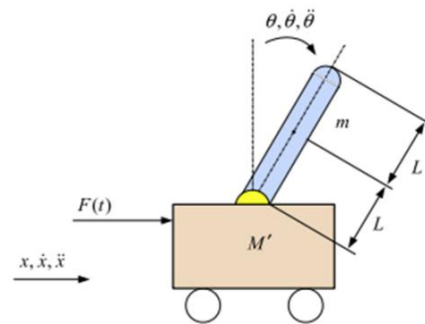


Figure 1: Inverted pendulum

With the notation x – cart position, θ – pendulum angle and F – applied force, the system can be described with the differential equations (1):

$$\begin{cases} \ddot{x} = \frac{1}{m+M} \cdot (F(t) + m \cdot l \cdot (\ddot{\theta} \cdot \cos(\theta) - \dot{\theta} \cdot \sin(\theta))) \\ \ddot{\theta} = \frac{m \cdot l}{I} \cdot (g \cdot \sin(\theta) - \ddot{x} \cdot \cos(\theta)) \end{cases} \quad (1)$$

As it is known, introducing variables $x_1 = x$, $x_2 = \dot{x}$, $x_3 = \theta$, $x_4 = \dot{\theta}$, we can reduce the problem to the standard form of a system of ordinary differential equations (2).

$$\dot{x}_i = f_i(x_1, x_2, \dots, x_n, t) \quad (2)$$

Obviously, to integrate the system, the dependence of the force F on time must be determined. So we need the simulation results to start the simulation process. Of course, this seeming paradox can be resolved by introducing mathematical description of the controller in the model. However, in most cases, simulation is used just in cases where the mathematical description of the complete control system is difficult to performer, so this theoretical approach is useless. It is important to note that the methods of numerical integration of systems of ordinary differential equations are highly developed, and at present a great progress has been made in the development of software that implements these methods.

Nowadays, freely available libraries of so-called solvers are available to developers, so it is highly desirable to use these components as part of modeling systems. Another significant problem is the way in which continuous signals are represented in implementation of modeling system. Most software systems and special-purpose languages designed for modeling dynamic systems use the discrete time model.

The process of simulation is to consistently calculate the state of the system at each time interval. Of course, this model of computation is not free from contradictions. If system is an aggregation of subsystems, subsystems may be connected in ways that yield the degree of ambiguity in computation. For example, assume that subsystem A has two outputs, one goes to subsystem B and another to subsystem C. Subsystem B has an output that feeds C. In this case, we may calculate the output of C whenever we have computed one of its inputs. Assuming that A has been processed, then we have the choice to calculate the outputs of B or of C. Depending on the choice of processing B or C, the outputs of C may have different values. Simultaneous events may in fact yield a nondeterministic behavior.

To reduce the inertia of our developed intelligent system for optimal energy-efficient control of the processes of the fuel-air mixture making in steam-driven boilers and to enhance its reliability, we suggest using several measured parameters of the system. Together with indirectly measured output power, it is assumed to take into account the concentration of carbon dioxide in the products of combustion as well as concentration of fuel dust in the mixture being made. The main idea of a proposed method is an active search of optimal composition at a given power. It is similar to gradient approach of searching the extremum of function of several variables used in numerical methods. Of course, hill-climbing technique, being applied in numerical analysis, cannot be directly applied to the problem considered due to a number of factors, the most important of which are the following two: first, the rate of convergence should be comparable to the rate of change of external factors (the given system operates in real time). Secondly, the value of deviating figure of generated power from the set value is strictly limited. To meet all the requirements it is assumed to use the vector of inputs of automated control system and to elaborate a mathematical model of external parameters influence on power output and measured parameters.

Basic research methods: mathematic modelling and live experiment: algorithm testing on model and experimental works on intelligent control system implementation, testing the model at actual data. An experimental stand will be built consisting of a small-sized boiler, burning nozzle, fuel intake system and necessary automation tools to construct a mathematical model, describing dependence of outputs on regulation characteristics (fuel consumption and concentration of air-fuel mixture) and external factors (coal grade, moisture content in the mixture, ambient temperature, etc.).

III. RESULTS AND DISCUSSIONS

We introduce experimental simulation software platform around the ecosystem Racket integrated with a domain-specific language (DSL) tool-chain for modeling of automatic control systems. Racket is a LISP-family general purpose programming language, as well as a platform for language creation and implementation. Racket ecosystem consists of implementation of Racket language itself (including run-time system, libraries, JIT – compiler) along with development environment called Dr. Racket.

The key point of a project is usage of metaprogramming and code generation techniques for implementation of described above. Powerful Racket macro system allows not only to expand the syntax of the core language, but also to perform some validation of code. The key features of the DSL are the separation of the description of the controller and the description of the plant, and use of the FRP (Functional Reactive Programming) technique to describe the controller.

We believe that the FRP approach is especially suitable for modeling automatic control systems for a number of reasons: firstly, a high level of abstraction allows developer describing the system ignoring the internal implementation of continuous and discrete signals and synchronization problems. Secondly, a high level of abstraction allows choosing different methods of signal representation for program implementation. In recent years, there has been a growing interest in the use of functional programming methods for modeling dynamic systems, and considerable progress has been made in this area [10,11]. The DSL expansion has much in common with the Haskell DSL YAMPA, in particular the use of so-called signal combinators.

To provide detailed training to the fundamentals of FRP, the language of YAMPA and its practical application to the modeling of dynamic systems, we recommend referring to the article [12-14], and below we give only a brief description of the main concepts of the language. The basic concept of functional reactive programming is that of a signal. Signal can represent any continuous, time-varying value. One can think of a signal as having polymorphic type (3):

$$\text{Signal } a = \text{Time} \rightarrow a \quad (3)$$

That is a value of type Signal A is a function that maps suitable values of time (double in most of cases) to a value of type A. It should be noted that type A is arbitrary, in particular, is a functional type. Thus, the signal is a high-level abstraction that does not always conform the intuitive notion of a signal as a time-varying value of a physical value. Nevertheless, the direct conforming between the physical equations (i.e. the specification) and the FRP code (i.e. the implementation) is exact.

A key feature of the YAMPA language that is avoidance of the signals as first-class objects usage contrasts YAMPA with other programming languages. In other words, programmer cannot access the value of a signal at given moment of time or create a signal. Instead, the programmer has an access only to signal convertors, or what we prefer to call signal functions. A signal function is just a function that maps signals to signals (4):

$$\text{SF } a \ b \rightarrow \text{Signal } a \rightarrow \text{Signal } b \quad (4)$$

However, the actual representation of the type SF in Yampa is hidden (i.e. SF is abstract), so one cannot directly plot signal functions or apply them to signals. Instead of allowing the user to define arbitrary signal functions from scratch (which makes it all too easy to introduce time- and space-leaks), we provide a set of primitive signal functions and a set of special composition operators (or “combinators”) that enables to

define more complex signal functions. These primitive values and combinators together provide a disciplined way to define signal functions that, fortuitously, avoids time- and space-leaks. Yampa program expresses the composition of a possibly large number of signal functions into a composite signal function that is then “run” at the top level by a suitable interpreter.

One of the significant differences between the DSL that we developed and YAMPA is the introduction of a signal source. As the source of the signal, interpolators are used, which “generate” a continuous signal by discrete values stored in a text file. Plant is modeled by a system of ordinary differential equations of the first order. The syntax of the controller description is as follows:

```
(plant < name of module > ((input signals)) ((output signals)) func - list)
(((internal signal description)) (map func - list signal - list))
```

Semantically plant definition is a closure, returning signal function, which arguments are input signals, and the result of conversion are output signals. Parameter funk-list is a list of functions (f₁, f₂, ..., f_n) representing the right part of ODE system in form (2)

IV. CONCLUSION

Automated control systems modeling is an area where the methods of declarative languages enable the technic to advance. The choice of the method of representing continuous signals is extremely important in the design of modeling systems as a whole. When modeling hybrid systems, in particular automatic control systems, it is desirable to be able to mix different representations of the signal during the simulation process.

The use of the FRP technique, in which the signals are first-class objects, allows great flexibility in the choice of the representation of continuous signals. Equally important are the advantages of FRP in the modeling of hybrid systems [15].

Although we have not completed an implementation of experimental software platform, this paper demonstrates our basic design approach and maps out the design landscape. We expect that further research of the links between functional reactive programming and automated control systems modeling will produce significant advances in this field.

The practical implementation of our developed software platform to design the intelligent system for optimal energy-efficient control of the air-fuel mixture making in steam – driven boilers can increase the efficiency of intake fuel control in steam boiler units will allow increasing the efficiency of steam boiler units, reducing the amount of chemical and mechanical fuel underburning, and, consequently, reducing the amount of combustion products, polluting the atmosphere.

ACKNOWLEDGMENT

The study was conducted with the financial support of the Science Committee of MES of RK in the framework of the program target financing for the 2017 biennium by the program 0006/PTF-17

REFERENCES

- [1] Ya.L. Packer, V.E. Egorov, “Investigation and analysis of dust incineration, taking into account the unevenness of processes over time,” *Teploenergetika*, no. 6, pp. 49-52, 1971. (In Russian).
- [2] M Kuang, Z. Li, C. Liu, and Q Zhu, “Experimental study on combustion and NOx emissions for a down-fired supercritical boiler with multiple-injection multiple-staging technology without overfire air,” *Applied Energy*, vol. 106, pp. 254-261, 2013.
- [3] M.Ya. Hesin *Indicators of combustion process for automation of combustion*. Moscow: Energy, 1969, pp.62. (In Russian)
- [4] A. Mallik, “State feedback based control of air-fuel-ratio using two wide-band oxygen sensors,” *ASCC, Malaysia*, pp. 1-6 [Proceedings of 10th Asian Control Conference Malaysia, p. 1-6, 2015].
- [5] G. Conte, M. Cesaretti and D. Scaradozzi, “Combustion control in domestic boilers using an oxygen sensor,” *IEEE Control and Automation (MED '06), Italy* [14th Mediterranean Conf., p. 1-4, 2006].
- [6] P. Neuman, B. Sule, P. Zitek. and T. Dlouhy, “Non-linear engineering simulator of a coal fired steam boiler applied to fault detection of optimum combustion control,” *IFAC Conference on Fault detection, supervision and safety for technical processes, Hungary*, p. 905-910, 2000.
- [7] W. Tan, H.J. Marquez and T. Chen, “Multivariable Robust Controller Design for a Boiler System,” *IEEE Trans. on Control Systems Technology*, vol.10, no5, pp. 735-742, Sept. 2002.
- [8] S. Simani, S. Beghelli, “PID controller design application based on boiler process model identification,” *46th IEEE Conference, USA*, pp. 1064-1069, December 2007.
- [9] J. Pieter Mosterman, “An overview of hybrid simulation phenomena and their support by simulation packages,” in *Lecture Notes in Computer Science*, number 1569, Fritz W. Vaadrager and Jan H. van Schuppen Eds. Hybrid Systems: Computation and Control '99, 1999, pp. 165-177.
- [10] M. Moser, M Pfeiffer., J. Pichler, “Domain-specific modeling in industrial automation: Challenges and experiences,” *Proceedings of the first International Workshop on Modern Software Engineering Methods for Industrial Automation*, pp. 42-51, 2014.
- [11] V. Djukic, A. Popovic, J.P. Tolvanen, “Domain-Specific Modeling for Robotics: From Language Construction to Ready-made Controllers and End-user Applications,” *Proceedings of the third Workshop on Model-Driven Robot Software Engineering*, pp. 47-54, 2016.
- [12] Z. Wan and P. Hudak, “Functional reactive programming from first principles,” in *Proceedings of the ACM SIGPLAN, USA*, pp. 242-252, 2000 [Conference on Programming language design and implementation, PLDI '00, USA, pp. 242-252, 2000] doi: <http://doi.acm.org/10.1145/349299.349331>.
- [13] H. Nilsson, A. Courtney, and J. Peterson, “Functional reactive programming, continued,” in *Proceedings of the 2002 ACM SIGPLAN workshop on Haskell, USA*, pp. 51-64, 2002] doi: <http://doi.acm.org/10.1145/581690.581695>.
- [14] P. Hudak, A. Courtney, H. Nilsson, and J. Peterson, “Arrows, robots, and functional reactive programming,” in *Advanced Functional Programming, Vol. 2638 of Lecture Notes in Computer Science*, J. Jeuring and S. Jones, Eds. Springer: Berlin / Heidelberg, 2003, pp. 1949-1955. URL http://dx.doi.org/10.1007/978-3-540-44833-4_6. 10.1007/978-3-540-44833-4_6.
- [15] H. Nilsson, J. Peterson, P. Hudak, “Functional hybrid modeling,” in *Proceedings of PADL'03, USA*, vol. 2562, pp. 376-390, January2003 [5th International Workshop on Practical Aspects of Declarative Languages, USA, p. 376-390, 2003].

Performance Comparison of Different Classifiers for Hungarian Handwriting Recognition

Gaye Ediboğlu Bartos*, Yasar Hoşcan**, Éva Hajnal*

* Anadolu University/Department of Computer Engineering, Eskişehir, Turkey

* Bilecik Şeyh Edebali University/Department of Computer Engineering, Bilecik, Turkey

** Anadolu University/Department of Computer Engineering, Eskişehir, Turkey

*Óbuda University Alba Regia Technical Faculty, Székesfehérvár, Hungary

gayeediboglu@anadolu.edu.tr, gaye.ediboglu@bilecik.edu.tr, hoscan@anadolu.edu.tr, hajnal.eva@amk.uni-obuda.hu

Abstract— Offline cursive handwriting recognition is an ongoing challenge due to the different styles used by different persons. The difference in the handwriting styles brings about the hardship for segmentation of the characters hence the overall accuracy of the recognizer is highly dependent on the style. In Hungary, there is a tradition of using cursive handwriting and the alphabet contains some letters with punctuation. Therefore, Hungarian handwriting recognition is a challenging task to perform. In this study, we compare the performance of different classifiers on a small data set (1750 characters, 50 samples for each letter in the alphabet) that has previously been generated for our study. The data set only consists of lower case Hungarian letters (35 letters excluding the ones which consist of two letters). In our study we compared the performance of four classifiers namely, Neural Networks, Support Vector Machines (SVM), Rough Sets Theory (RST) and Bayesian Networks (BN) using the WEKA machine learning tool. The results indicated that in terms of classification accuracy, neural networks performed the best followed by BN, SVM and RST respectively. However, in terms of the time taken to build the model neural networks performed the poorest. BN took the shortest time to build followed by SVM and RST respectively.

I. INTRODUCTION

Optical Character Recognition (OCR) is conversion of scanned images of machine printed or handwritten text, numerals, letters and symbols into a computer processable format such as ASCII without any human intervention. There are two types of OCR namely online and offline recognition. In online recognition, the characters are recognized as they are drawn. Furthermore, the order of strokes are available and successive points are represented as a function of time [1][2]. On the other hand, in offline recognition optical recognition is performed after the writing or printing has been completed. In other words, its input is an image or a scanned document [3].

An OCR system consists of several components. Fig. 1 shows the components in a typical OCR system. As can be seen from the Fig. 1, firstly the document is scanned through an optical scanner. Secondly the crucial pre-processing phase is applied. Pre-processing is critical for an OCR system since the outcomes of this step are going to be recognized in the next step. Generally in the pre-processing phase binarization, noise removal, normalization, feature extraction and segmentation are performed. Finally in classification step, the recognition is performed. In addition to those steps, an extra post-

processing phase could be adopted in which verification is performed in order to improve the accuracy rate.



Figure 1 Components of a typical OCR system

This paper compares the performance of four classifiers applied to a small dataset which was created by the researchers earlier. The classifiers adopted are Neural Networks, Support Vector Machines (SVM), Rough Sets Theory (RST) and Bayesian Networks (BN). The next section provides the properties of Hungarian Handwriting with reference to its challenges. In the following sections the adopted dataset, feature extraction, classification phases are explained and the results are provided. Finally the conclusion is presented.

A. Properties of Hungarian Handwriting

Hungarian Language consists of 44 letters (Fig. 1). Some Hungarian letters are the same as English letters, however other letters have punctuation and some consist of more than one letter. These characters of the language generate a challenge for recognition purpose such as removal of the punctuation at the noise removal phase.

Another challenge in recognizing Hungarian handwriting is that in Hungary there is a tradition of using cursive scripts. Cursive character of the handwriting brings about the challenge to the segmentation phase. However, this study does not include the segmentation of the Hungarian handwriting. The dataset adopted is already segmented into the characters. However, due to the nature of cursive handwritings, the characters are not as readable as in discretely written texts. The characters may be distorted and written in a personal way which is not clearly readable. In addition to the challenges, there are not many studies conducted for the purpose of Hungarian Handwriting Recognition.

a á b c cs d dz A Á B C Cs D Dz
 dzs e é f g gy h Dzs É Ê F Gy Gy H
 i í j k l ly m n J J J K L Ly M N
 ny o ó ö p q r Ny O O O P Q R
 s sz t ty u ú ü ü S Sz T Ty U U U U
 v w x y z zs U W X Y Z Zs

Figure 2 Hungarian Alphabet [4]

II. DATA SET

The adopted dataset was previously created by the researchers. It includes 1750 characters (50 samples of 35 lower case Hungarian characters excluding the characters which consist of more than one letter). Each character in the data set is normalized to 28x28 pixels and in the skeleton form as can be seen in the Figure 3. These characters are the output of the previous stages of study and in this study they are used as input of classification.



Figure 3 Sample characters from the dataset

In order to create the data set, Hungarian handwritings were collected from multiple users on paper. Then the papers were scanned with 300 dpi and saved in the png format in order to avoid information loss. Consecutively, the documents were pre-processed. Pre-processing phase included binarization, skew correction, slant removal and noise removal. Thereafter, the lines, words and characters were segmented respectively. Finally, size of segmented characters was normalized into 28x28 pixels.

III. FEATURE EXTRACTION

In the feature extraction phase, significant features of a character are extracted. The result of the classification is directly affected by the features extracted since the feature vectors are going to be the input for the classifier. Therefore, it is crucial to extract the key features. It is possible to group the features into three categories namely distribution of points, structural analysis and transformations and series expansions. In our work, features were extracted using distribution of points and structural analysis features.

A. Distribution of Points:

In this category, features are extracted based on the statistical distribution of points. These features are usually tolerant to distortions and style variations[5]. The feature extraction techniques used in this study which are based on distribution of points are represented below:

Projection Profiles: Profiles refer to the distance from the border of the image until the next white pixel. An example representation of projections of a character is given in the Figure 4. In our work, left, right, top and bottom profiles are used as feature vectors.



Figure 4 Right, left, top and bottom profiles of a character

Extremas of the character image: It returns the x and y coordinates of the 8 extremas of the image namely top-left, top-right, right-top, right-bottom, bottom-right, bottom-left, left-bottom and left-top as can be seen in Figure 5.

bottom-left, left-bottom and left-top as can be seen in Figure 5.

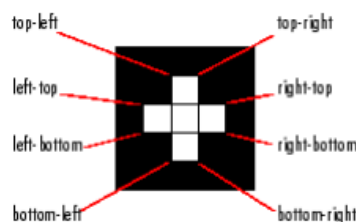


Figure 5 Extremas

Center of gravity: The (x, y) values of the center of gravity of the character image. In addition to those, the distance between the bottom of the character and the y coordinate of the center of gravity and the distance from the left end of the character and x coordinate of the center of gravity are also used as feature vectors.

Density: The density of the character image.

Area: Actual number of pixels in the region, returned as a scalar.

The proportion of width and length: The result of dividing the width of the character into the length of the character.

Number of regional minimas and maximas: After applying horizontal and vertical projections, the number of regional maximas and regional minimas for both vertical and horizontal projection are used as features.

B. Structural analysis

This type of features represents the geometric and topological structures of a character. The most common types include endpoints, loops and strokes[5][6]. It is worth mentioning that these types of features are highly affected by any noise in the data. As can be seen in Table 1 that any noise in the character image would cause a change in the feature vector thus it is crucial for the recognition that the data set is noise free. The feature extraction techniques used in this study which are based on the structural analysis are explained below:

Table 1 An example feature set of noise free character image and a noisy character image

Character image	#endpoints	#connected components	#isolated small areas
	4	2	2
	5	3	3

No of endpoints: It represents the number of pixels having only 1 connected neighbor in an 8 connected image.

No of branch points: It represents the number of pixels having at least 3neighbors that are 1s in an 8 connected image.

Euler number: The Euler number represents the total number of objects in the image minus the total number of holes in those objects.

Number of loops: The vector represents the number of holes in the image.

Number of small components: The feature vector represents the number of isolated areas with the area smaller than 7 in the character image.

Sum of the area of small components: It represents the sum of all small components with less than 7 pixel area.

No of connected components: It represents the number of connected components in the image.

IV. CLASSIFICATION

The data was classified by four classifiers namely, Neural Networks, Support Vector Machines, Rough Sets Theory and Bayesian Networks using WEKA machine learning tool[15]. A brief explanation of the classifiers is given below.

Neural Networks: Neural Network design which is made of parallel interconnection of adaptive processors[7]. Since it has the parallel connections, it has a better performance than the classical techniques. Additionally, its adaptive nature provides a better adaptability to changes in the data and an ease to learn the characteristics of input signal[8]. The structure of a neural network contains many nodes. The output of one node is input to another, thus the final output is a result of complex interaction of all nodes. Neural network architectures can be classified into two major groups which are feed-forward and feedback networks. In our work, the multilayer perceptron of the feed forward networks is adopted since it is the most popular for character recognition purposes.

Support Vector Machines: SVM classifier carries out the classification by mapping all the input data to a value in a higher dimensional space. The data is classified by coming up with an optimum N-dimensional hyper plane which separates data into positive and negative examples[9]. In our work SVM is used due to its ease of implementation and high performance.

Rough Sets Theory: Rough Sets is a mathematical tool which deals with uncertainty and vagueness[10]. The idea is based on the assumption that with every object of the universe of discourse, it is possible to associate some information. Objects characterized by the same information are indiscernible in view of the available information about them. The indiscernibility relation generated in this way forms the mathematical basis of the theory. The rough sets theory provides a technique of reasoning from imprecise data, discovering relationships in data and generating decision rules[11]. Not requiring any preliminary or additional information about data like probability distributions in statistics is rough sets theory's main strength[12]. In our work, the RST was adopted due to the above-mentioned strengths. Since the data set is relatively small, we believe that RST may be a good classifier in such conditions.

Bayesian Networks: Bayesian classifiers are the statistical classifiers based on Bayes Theorem. They are

able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class[13]. Bayesian networks are a model representing uncertain knowledge about a complex phenomenon and allowing real reasoning from data. They effectively represent a domain of knowledge, as a causal graph, permitting learning the dependency relationships that can help us make decisions and manage all incomplete data[14].

V. EXPERIMENTS AND RESULTS

The classification task was carried out with and without applying feature selection. For the same data set, a supervised feature selection algorithm provided by WEKA is applied to the features. Additionally, the results of the classification without any feature selection are also given. Finally the recognition is performed with three different cross validation values which are 5, 7 and 10 fold cross validation.

The classification accuracy and time taken to build the model for each classifier are given in the Table 2 and Table 3 respectively.

Table 2 Classification accuracies for different classifiers

	No Feature Selection			Feature Selection		
	5 fold	7 fold	10 fold	5 fold	7 fold	10 fold
SVM	91.1 %	92.5 %	92.1 %	95.0 %	95.4 %	95.6 %
RST	86.1 %	88.8 %	88.1 %	88.4 %	91.1 %	90.3 %
BN	89.0 %	88.8 %	89.4 %	95.8 %	96.0 %	95.4 %
NN	92.2 %	92.9 %	92.7 %	96.6 %	96.8 %	96.6 %

As provided in Table 2, Neural Networks give the highest accuracy with and without feature extraction compare to the other classifiers. It is followed by BN, SVM and RST respectively. In addition to that, it is possible to say that feature selection increases the accuracy as well as the time taken to build the model. Finally, a 7 fold cross validation appears to be the most suitable value for this data set since it is fastest and provides the most accurate classification results.

Table 3 Time taken to build the model (seconds)

	No Feature Selection			Feature Selection		
	5 fold	7 fold	10 fold	5 fold	7 fold	10 fold
SVM	1.81	0,84	0,8	0,69	0,69	0,71
RST	21,01	19,8	20,02	14,09	10,95	11,52
BN	0,11	0,06	0,09	0,2	0,06	0,03
NN	2301	2158	2265	1567	1452	1504

Although, NN gives the best accuracy, it is clearly the slowest when it comes to the time taken to build. There is almost 99% difference in speed with the second slowest classifier RST. Although there is only about 1%

difference in accuracy with the second best classifier, the time taken to build the model is almost 99% times slower.

VI. CONCLUSION

Hungarian Handwriting recognition is a challenging field considering the tradition of using cursive handwriting and the letters with punctuations. In this study, we performed a Hungarian Handwriting classification using a small data set with four different classifiers. The results of the different classifiers are compared in terms of their performances.

Classification of handwritten characters includes several crucial steps. It is possible to say that feature extraction is one of the most important steps for the recognition of the characters since the distinctive and characteristic features must be extracted. In our work, several feature extraction methods were adopted. Consecutively, the character images from the data set were classified by four different classifiers.

The results were interesting considering the difference in the time taken to build different classifiers. NN performed the best in terms of accuracy, followed by BN, SVM and RST. However, NN was significantly slower than any other classification with around 99% difference in speed with the second slowest RST.

VII. FUTURE WORK

A deeper understanding of the results of feature extraction methods may be useful with representation of data such as which method is more distinctive for which characters and which characters are more likely to be misclassified. Additionally, it would be beneficial to apply the same method to a bigger data set. We believe the greater the data set, the better the accuracies are going to be. For example, RST was applied considering its nature to work well with only a little data available. However, it performed one of the poorest in both accuracy and time taken to build the model. It would be necessary to compare the differences with a bigger data set.

VIII. REFERENCES

- [1] J. Pena, S. Letourneau, and F. Famili, "Application of Rough Sets Algorithms to Prediction of Aircraft Component Failure," in *Advances in Intelligent Data Analys*, no. i, 1999, pp. 473–484.
- [2] R. Plamondon and S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, 2000.
- [3] P. K. Charles, V. Harish, M. Swathi, and C. H. Deepthi, "A Review on the Various Techniques used for Optical Character Recognition," *Int. J. Eng. Res. Appl.*, vol. 2, no. 1, pp. 659–662, 2012.
- [4] Bergamott, "Írott és nyomtatott ábécé," 2010. .
- [5] L. Eikvil, "Optical character recognition," *Citeseer. Ist. Psu. Edu/142042. Html*, vol. 3, no. 1, pp. 4956–4958, 1993.
- [6] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on offline handwriting," *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.)*, vol. 31, no. 2, pp. 216–233, 2001.
- [7] R. K. Nath and M. Rastogi, "Improving Various Off-line Techniques used for Handwritten Character Recognition: a Review," *Int. J. Comput. Appl.*, vol. 49, no. 18, pp. 11–17, 2012.
- [8] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 31, no. 2, pp. 216–233, 2001.
- [9] J. Taylor, S. Kumar, and I. Khaimovich, "Cursive Handwriting Segmentation and Character Recognition," 2007.
- [10] Z. Pawlak, "Rough Sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.
- [11] J. J. Shuai and H. L. Li, "Using rough set and worst practice DEA in business failure prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3642 LNAI, pp. 503–510, 2005.
- [12] B. Mağden and S. Telçeken, "Probabilistic Rough Sets in Turkish Optical Character Recognition," in *6th World Conference on Soft Computing*, 2016, no. 3, pp. 170–173.
- [13] K. M. Lakshmi, K. Venkatesh, G. Sunaina, D. Sravani, and P. Dayakar, "Hand Written Telugu Character Recognition Using Bayesian Classifier," *Int. J. Eng. Technol.*, vol. 9, no. 3S, pp. 37–42, 2017.
- [14] K. Jayech, M. A. Mahjoub, and N. Ghanmi, "Application of Bayesian Networks for Pattern Recognition: Character Recognition Case," in *International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, 2012, vol. 3, no. March, pp. 748–757.
- [15] F. Eibe, M. A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", *Morgan Kaufmann*, Fourth Edition, 2016.

Creation of LED network of data transmission based on Visible Light Communication technology

A.Baklanov*, O. Baklanova*, K.Eleusizova*, V.Sayun**, E.Grigoryev**

* D.Serikbayev East Kazakhstan State Technical University/ Faculty of Information Technology and Energy, Ust-Kamenogorsk, Kazakhstan

** Tomsk State University of Control Systems and Radioelectronics/ Faculty of Electronic Engineering, Russia
ABaklanov@ektu.kz, svm@ie.tusur.ru

Abstract—One way to improve the characteristics of data transmission in wireless networks is to transition to the optical wavelength range. This became possible with the appearance of white LEDs used for lighting. Since 2011, the new technology Visible Light Communication (VLC) is rapidly developing - a technology that allows the light source, in addition to lighting, to transmit information using the same light signal. This technology can use LEDs at speeds up to 500 Mbps. In the article, the amplitude-frequency characteristics of light-emitting diodes of visible light, the principle of their action and technical characteristics necessary for constructing networks using VLC technology are studied.

I. INTRODUCTION

Currently Wi-Fi is the most widely used data communications technology, and using it helps setting up local computer networks and allows connecting and transmitting data to mobile devices. However, the technology has limitations on data transfer rate related to electromagnetic radiation wave length, and, based on medial research, emission intensity near a router assuming a large amount of users may have harmful effects on human health. One way to improve data transfer characteristics in wireless networks may be the conversion to optical wavelength range. This was made possible by the appearance of white LEDs used for lighting. Since 2011 a new technology of VLC (visible light communication) has been rapidly progressing, the technology allowing a light source to not only illuminate a room, but also transfer information using the exact same light signal [1]. VLC uses visible light in optical spectrum (about 400-800 THz). This technology may utilize fluorescent lamps for signaling at about 10 Kbit/s or LEDs for signaling at about 500 Mbit/s. This project proposes developing and creating a prototype of a new generation data transfer wireless network Li-Fi (Light Fidelity or Light-based Wi-Fi) based on an LED system used for illuminating a room.

Since 2011 Harald Haas, optical wireless data transmission specialist and a Professor at the University of Edinburgh (Edinburgh, the United Kingdom), was seriously advancing the new technology of wireless data transmission through blinking light-emitting diode [2,3]. At that time the majority of university professors decided that the idea was definitely interesting, but hardly

implementable. Four years later Haas has created the first router that works according to his conception.

The technology was called Li-Fi. The new router showed amazing capabilities. It surpassed Wi-Fi in speed 100 times. The new router achieved record data transmission at 224 Gb/s in the laboratory conditions. The test was performed by the Estonian company Velmenni in the laboratory. Haas provided his first router with a solar cell battery to make the network access offline. Currently the router has stable data transfer rate at 10 Gb/s through barely noticeable blinking LED [4].

In order to deliver the first serial systems to the European market the Li-Fi inventor Harald Haas consolidated his purecompany with Lucibel company to collectively develop and effectively advance the innovation closer to an average consumer in order to make Li-Fi the main way to access the network for users.

The core of technology works according to the following scheme. Three color channels of the miniature LED lamp (red, green, and blue) transmit data in parallel up to 3.5Gb/s. As the result we can obtain 10Gb/s. Turning on or off the light occurs at breakneck speed that creates enormous aggregation of binary data.

This is called digital modulation with orthogonal frequency-division multiplexing (OFDM), and it allows transmitting millions of light beams with different intensity per second.

Professor Haas demonstrates it with shower head example that spouts strictly in parallel, the light in Li-Fi system working much in the same way.

Meanwhile Chinese and German researches took an interest in researching this topic. As far back as in 2011 the Germans could achieve data transmitting with record rate 800Mb/s at 1.8m distance, and the Chinese connected 4 computers to the internet at 150 Mb/s speed rate.

Professor Haas accentuated that the light waves technology is more reliable in terms of security than Wi-Fi. It is known that it is easy to hack into the Wi-Fi network from outside and intercept the files, since the radio waves pass through the walls beyond premises.

In the meantime the Li-Fi traffic can theoretically be captured only if you are in the same room, where the transmitter and receiver are located, since the light can't pass through the walls. Thus a reliable barrier is set up for the intruders, they won't be able to hack or intercept anything either from a street or even from the next room.

But first and foremost the advantage of Li-Fi is in the high speed rate and low power consumption (the standard routers' efficiency reaches 5% in the best case).

Definitely there are future prospects for the technology. The visible light waves have very wide frequency band, it is 4 times wider than the radio waves. There is no risk that the networks become overloaded, it won't lose either speed rate or the network performance like with Wi-Fi [5].

The LEDs are widespread. The infrastructure is almost here, and in addition the LEDs can fulfill dual roles - data transmitter and source of light at the same time. But there is still a question, how correct will the work of the system be in the illuminated room or in the bright sunlight condition.

Yet the Developers have high hopes for VLC - for visible light data transmission that is how this technology is called in scientific terms [6].

The high speed rate of Li-Fi already allows to successfully transmit video streams in HD quality, while keeping up high power performance of the system [7]. Another advantage over Wi-Fi is the accuracy and stability with the internet connection inside the buildings. The weak and intermittent signal area problem is solved due to the equalized LED transmitters distribution [8].

II. REALLY MODEL OF ELECTRONIC DEVICES

The structural scheme of data transmission using an LED lighting device is shown in Fig. 1.

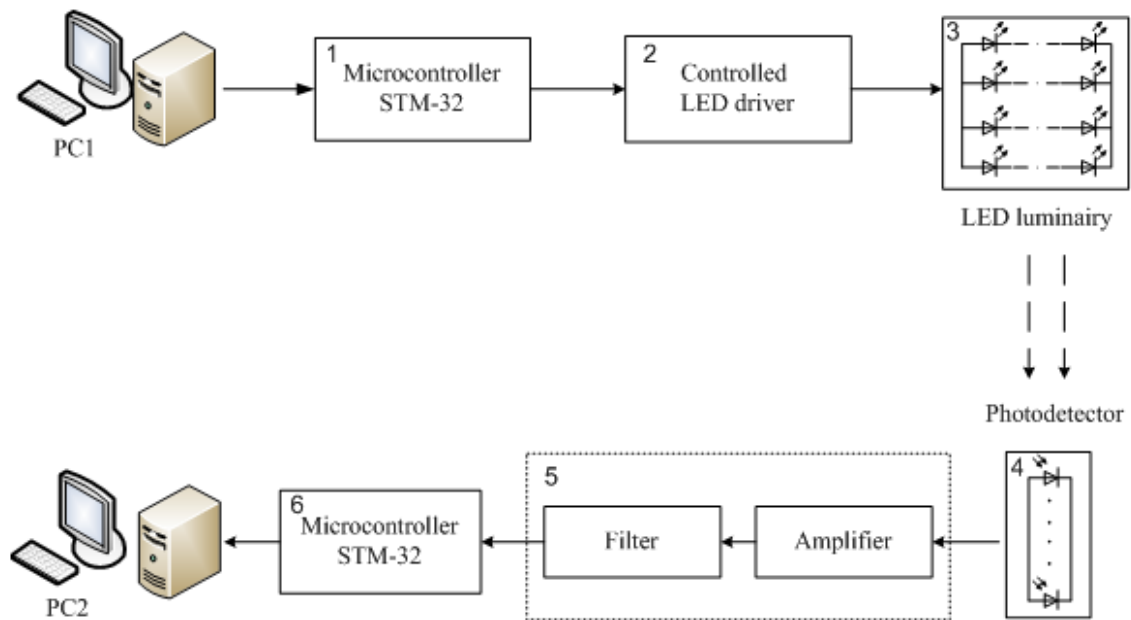


Figure 1. Structural scheme of data transmission system on VLC technology

The transmitting unit of the system consists of three parts. Block (1) - preparing data in a specific format. We use the microcontroller STM 32, which allows to match the signal coming from the personal computer (PC) with the lighting control system.

Block (2) - lighting control. This can be a standard driver, built on the basis of the MAX16800 or LM3404HV, which supports the dimming mode. You can use the control scheme shown in Figure 2.

Block (3) - LED luminaire. It consists of 4 LED strips, each of which includes 8 Nichia LEDs connected in series with 1W power.

The receiving unit also consists of three parts. Photodetector (4) based on the photodiode array. Further, a signal amplifier with a high-pass and low-pass filter and a matching device (5). We also use the STM32 microcontroller (6).

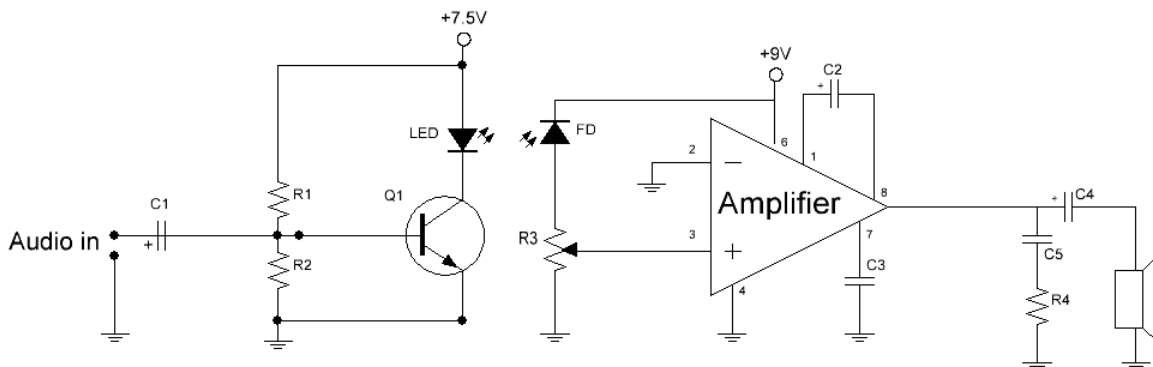


Figure 2. Schematic scheme of the device for transmitting sound with a white LED

To demonstrate the work and test this scheme we created model, a circuit diagram of which is shown in Fig.2.

In this circuit, to control LED, we used an amplifier assembled on transistor Q1, providing the necessary current for the LED. We worked with transistor BC337 (originally with transistors 2N2222A, 2N4401).

For the transistor BC337, the nominal values of the circuit elements: $C1 = 2.2\mu\text{F}$, $R1 = 4.7\text{k}\Omega$, $R2 = 1\text{k}\Omega$. The LED was used with a nominal value of 1W produced by Nichia NCSL219B.

The authors have sufficiently well studied the lighting characteristics of this LED. LED was used in the design of luminaires and installation of LED systems for office lighting [9].

The investigation of the frequency characteristics of this LED showed (Fig. 3) that for a signal frequency of 250 kHz the signal does not change in amplitude. Further, a decrease from the original signal to a level of 3dB at a frequency of 3.4 MHz is observed. This means that data transfer using LED in lighting systems will be limited to a speed of 3.2 Mbit/s.

The photodiode BPW34S was used in the receiving path. The amplifier is assembled on the LM386 chip. The nominal elements of this amplifier: $C2 = 10\mu\text{F}$, $C3 = 0.1\mu\text{F}$, $C4 = 250\mu\text{F}$, $C2 = 0,05\mu\text{F}$, $R3 = 10\text{k}\Omega$, $R4 = 10\text{k}\Omega$.

For normal operation of the described scheme, it is necessary to conduct studies in the room without daylight, in order to remove the power supplies of the lighting devices. Also, in bright sunlight, data transmission was not possible.

If the input and output of this circuit are connected to microcontrollers, as shown in Fig.1, then real-time audio is transmitted from PC1 to PC2.

Microcontroller STM32F4 Discovery used to organize the transfer of sound. The USB input was used for the transmission channel. A signal from which it was converted using a 24-bit DAC and entered the input of the LED control circuit. An input-output port was used for the receiving channel. The signal to the port came from the amplifier after the low-pass filter C4. Then the signal was digitized and transmitted via USB to a personal computer.

Structural scheme of the STM32F4 Discovery card is shown in Fig. 4.

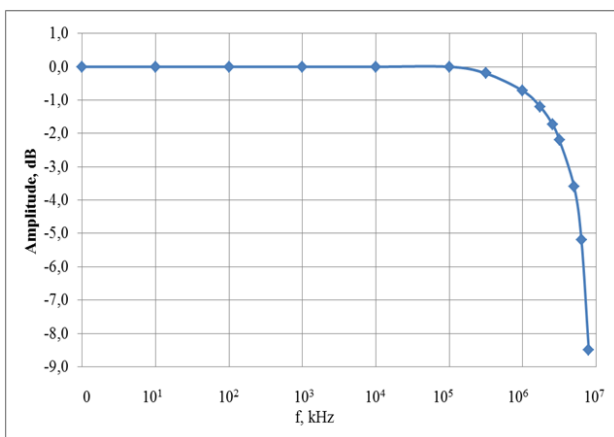


Figure 3. Amplitude-frequency characteristic of a white LED

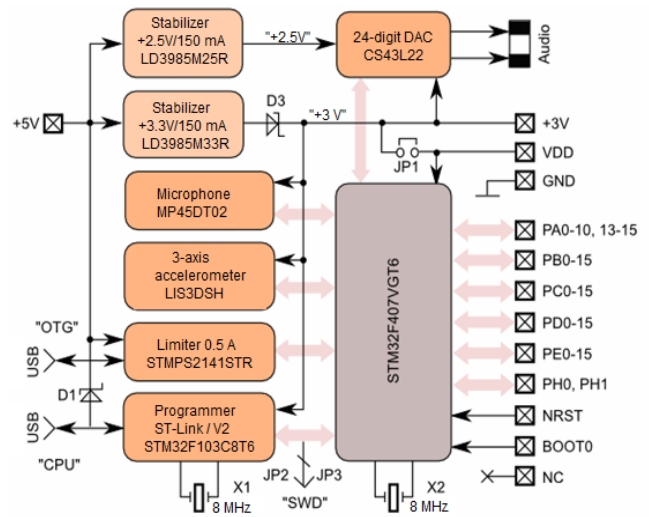


Figure 4. Structural scheme STM32F4 Discovery

The choice of this board for the experiment was due to the fact that this assembly has all the necessary components for working with sound – Analog I / O port, 24-bit DAC connected to the audio connector, USB port for data exchange with the computer. In addition, this board has a low cost (less than \$ 10). The appearance of the STM32F4 Discovery card is shown in Fig. 5.

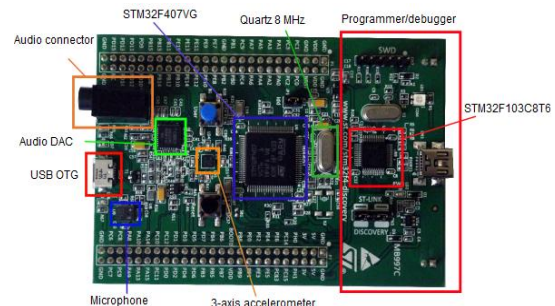


Figure 5. The appearance of the STM32F4 Discovery board

The microcontroller control program is standard. An example is the implementation carried out in [10, 11]. Distortions were almost absent when playing audio using microcontrollers in real time. However, detailed studies of the frequency characteristics of the transceiver-based communication channel based on white light, set forth in this work, were not carried out.

The appearance of the device assembled according to the structural scheme is shown in Fig. 6. In addition, headphones were used to control the sound quality, and a notebook was used to analyze the frequency characteristics of the LEDs, to program the microcontroller and to monitor the operation of the entire circuit.

The optimization of the placement of luminaries in the room plays an important role for the stable transmission of data using LED lighting devices. The decision of the task of choosing lighting devices and the optimization of their placement for uniform illumination of the room while performing the required level of illumination of the working surface of a particular room were published authors in [12, 13].

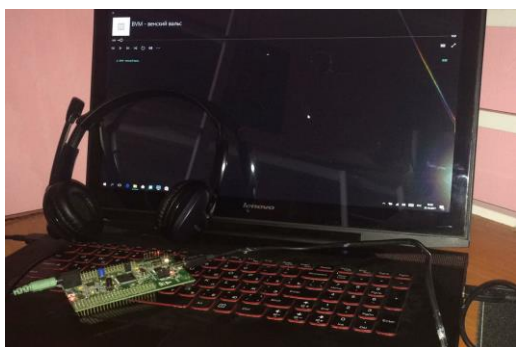


Figure 6. The appearance of the device

Calculations were carried out in DIALux and Statistica programs, simulation of the distribution of illumination was carried out in the MATLAB program.

An example of the optimal placement of LED lamps for a particular office space is shown in Fig. 7. Optimal arrangement of lamps is not accidental. LED luminaires 1, 2, 8 and 9 have the highest luminous flux (3120 lm) and are used as basic lighting devices. LED lamps 3, 4, 6, 7 have the smallest light flux (1980 lm) and are used to illuminate unreached "dark" areas of the room. The value of the light flux of the LED luminaire 5 (2390 lm) is the average between the values of other luminaires and is the central lighting device.

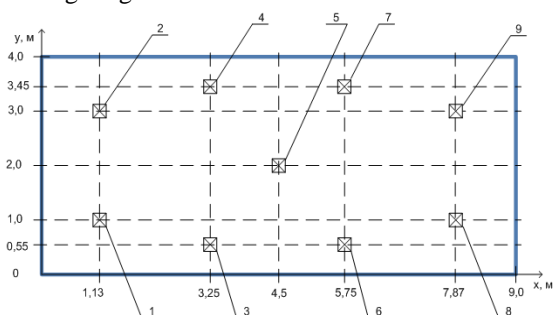


Figure 7.- The placement of LED lamps

After determining the location of the fixtures, a room illumination map was drawn up at the level of the working plane and a model of the distribution of illumination on the working plane was constructed (Fig. 8). It can be seen that the oscillation of the light intensity on the working surface does not exceed 10%, which allows to say about uniform illumination of the room.

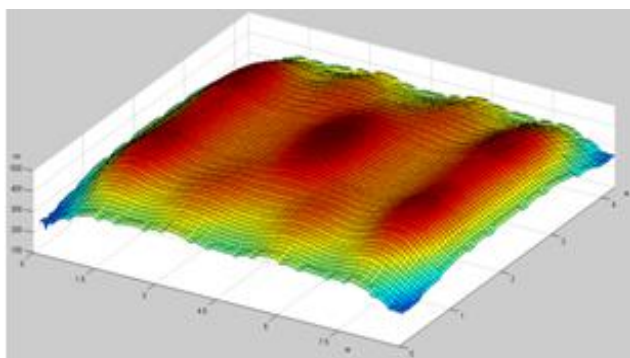


Figure 8. 3D model of the distribution of illumination on the working plane of the room

This arrangement allows for a stable reception of the signal at any point in the room.

CONCLUSION

The work demonstrates the transfer of data (sound) from one computer to another using VLC technology, while the implementation of the transfer is carried out using a simple scheme and a common element base.

To increase the data transfer rate, it is necessary to use the modulation of each color of the white LED. In this case, the hardware that supports higher frequencies is required as the transmitting receiving network. So for high frequencies of the order of 100 MHz the sensitivity of the LED falls by 12-13 dB. This circumstance indicates that for data transmission it is necessary to use more powerful LEDs or several LEDs in parallel.

ACKNOWLEDGEMENT

The study has been conducted with the financial support of the Science Committee of RK MES in the framework of the program target financing for the 2017-2019 biennium by the program 0006/PTF-17 "Production of titanium products for further use in medicine".

REFERENCES

- [1] Kumar N., Terra D., Lourenço N., Alves L.N., and Aguiar R.L. Visible light communication for intelligent transportation in road safety applications // in Proc. 7th Int. Wireless Commun. Mobile Comput. Conf. 2011. – pp. 1513 – 1518.
- [2] Ravi Prakash, Prachi Agarwal. The New Era of Transmission and Communication Technology: Li-Fi (Light Fidelity) LED & TED Based Approach // International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 3, Issue 2, February 2014.
- [3] https://www.ted.com/talks/harald_haas_wireless_data_from_every_light_bulb/transcript
- [4] https://www.ted.com/talks/harald_haas_a_breakthrough_new_kind_of_wireless_internet/transcript
- [5] Jitender Singh, Vikash. A New Era in Wireless Technology using Light-Fidelity // International Journal of Recent Development in Engineering and Technology, Volume 2, Issue 6, June 2014.
- [6] Karthika R., Balakrishnan S. Wireless Communication using Li-Fi Technology // SSRG International Journal of Electronics and Communication Engineering (SSRG-IJECE), volume 2, Issue 3, March 2015.
- [7] Dinesh Khandal, Sakshi Jain. Li-Fi (Light Fidelity): The Future Technology in Wireless Communication // International Journal of Information & Computation Technology, Vol.4, N 16, 2014.
- [8] Shubham Chatterjee, Shalabh Agarwal, Asoke Nath. Scope and Challenges in Light Fidelity(LiFi) // Technology in Wireless Data Communication International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 6, Volume 2, June 2015.
- [9] Grigoryeva S., Baklanov A., Gyorok Gy. Control of LED Lighting Equipment with Robustness Elements // Acta Polytechnica Hungarica. – Budapest.2016. – v.13, № 5. – pp.105–119.
- [10] Playing sound on STM32-Discovery with Speex. <http://we.easyelectronics.ru/STM32/vosproizvedenie-zvuka-na-stm32-discovery-pri-pomoschi-speex.html>
- [11] Playing sound on STM32F4 Discovery. <http://microtechnics.ru/vosproizvedenie-zvuka.html>
- [12] Grigoryeva S., Grigoryev E. Optimization of the placement of LED lighting for the organization of uniform illumination of office premises // Student: science, profession, life/ OSGUP. - Omsk, 2015. - pp.86-90.

Grigoryeva S., Baklanov A., Kvasov A. Modeling and experimental study of the placement of LED lighting fixtures for office space // Vestnik EKSTU. - 2015. - No. 3 (69). - pp.114-121.

Combined use of Sentinel-1A and Sentinel-1B data for determination of post-seismic vertical surface deformation - a case study: Perugia, 2016

Gergely LÁSZLÓ*, Lóránt FÖLDVÁRY*,**

* Óbuda University, Alba Regia Technical Faculty, Institute of Geoinformatics, Székesfehérvár, HUNGARY

** Hungarian Academy of Science, Research Centre for Astronomy and Earth Science, Geodetic and Geophysical Institute, Sopron, HUNGARY

laszlo.gergely@amk.uni-obuda.hu, foldvary.lorant@amk.uni-obuda.hu

Abstract— In this paper the feasibility of combined use of Sentinel-1 and Sentinel-1B images has been investigated by applying them for detecting vertical surface deformation due to a major earthquake event. The test event is the 24 August, 2016 event with Richter amplitude 6.2 in the vicinity of Perugia. The use of the two different satellites with identical technical parameters may be beneficial in the temporal resolution of the derived images. Even though in this study a successful application for the combination of the two independent images for InSAR processing has been demonstrated, it is just a case study, for a final proof more elaborated investigations should be performed.

I. INTRODUCTION

In April 03, 2014 a new era has begun in space geodesy and remote sensing with the launch of the Sentinel-1A satellite [1] starting a dedicated joint Copernicus program of European Space Agency (ESA) and European Union (EU). The initial mission plans have contained three Sentinel series (1-3) with three satellite each (A, B, C), but afterwards a new satellite (Sentinel-1D) and three more series (4-6) has been added to the mission, which are scheduled to be launch after 2020 [2].

In this paper measurements of Sentinel-1A and Sentinel-1B have been used. The main technical parameters are listed in Table 1 based on [1].

Apogee:	693 km
Orbit regime:	SSO (Sun-synchronous Orbit)
Inclination	98.2°
Period:	98.6 minutes
Repeat interval:	12 days
Mission duration:	7 years planned, 12 years of consumables
Mission objectives:	atmosphere-, marine- and land-monitoring, climate change, emergency management and security.
Launch:	03-04-2014 (1A) and 25-04-2016 (1B) from Guyana Space Center.

Table 1. Sentinel-1A and B technical parameters [1]

Using the state-of-the-art vertical monitoring method, the Synthetic Aperture Radar Interferometry (InSAR) technology [3], a wide area of applications ranging from

monitoring vertical displacement of discrete points, through monitoring movements of buildings, to observation of crustal movements on continental scale becomes feasible [4]. The SAR technology is an active remote sensing method [5], and as such, it is independent from the daytime cycle. The satellite radar interferometry is based on comparing one or more radar image. Using the phase values of the images, phase differences can be derived, which are then interfered at the steady (i.e. being in no motion) areas. Based on the interferences, an interferogram can be deduced to estimate the amount of vertical movement at the area of interest [6].

II. CASE STUDY: PERUGIAN EARTHQUAKE

The source of the Italian earthquakes is the subduction of the African plate under the Eurasian plate. These two plates are converging every year 2 cm to each other, causing smaller earthquakes almost constantly in the Italian peninsula. This motion has created the Alpine mountain range in the past, and will result in a total merge of the two continents in the future, diminishing the Mediterranean Sea eventually.

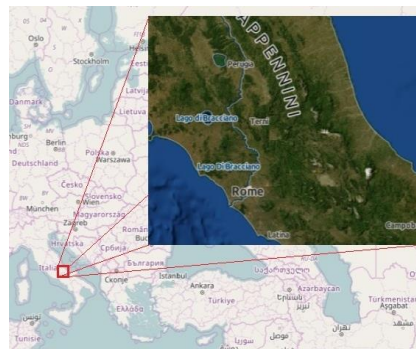


Figure 1. Location of Perugia

On 2016-08-24 an earthquake magnitude of 6.2 on Richter-scale shook the area of Perugia, 140 km North-East from Rome, c.f. Figure 1. Actually, several tremors been occurred within few hours as it is indicated by the purple circles on Figure 2, displaying seismic events according to the event catalog of the IRIS (Incorporated Research Institutions for Seismology) on that day [7]. Around Perugia as a consequence of these quakes some faults have appeared, manifesting that the area built on a very active part of Italy. [8]



Figure 2. The effects of the Perugian earthquake [7]

Sensing period:	2016/08/21-2016/08/27
Satellite Platform:	S1A_* or S1B_*
Product Type:	SLC
Polarization:	VV
Relative Orbit Number:	117

Table 2. Search parameters

Satellite	Orbit	Date	Time
S1A	desc.	2016-08-27	17:05:42.593
S1B	desc.	2016-08-27	05:10:35.191
S1A	asc.	2016-08-26	05:19:22.305
S1B	asc.	2016-08-21	17:05:09.646
S1A	desc.	2016-08-21	05:11:16.928

Table 3. Search results

III. DATA PROCESSING

The first step before any actual processing is to acquire the data to work with. The source of data of this study is the Sentinel Scientific Data Hub site, operated by ESA (c.f. Figure 3). The processing software is also provided by ESA developed by Brockmann Consult, Array Systems Computing and C-S. This software is designed to process images captured by any Sentinel satellites. [9]

During the selection of the proper images the first step is to decide on the use of data type, SLC or GRD data. The SLC contains phase and amplitude values, while the GRD contains amplitude only. As for InSAR deformation monitoring phase information is also required, the SLC data were used.

Before the launch of Sentinel-1B it was essential to use images 12 days apart from each other, since this is the time period, when the satellite returns to the same position in every 175 full orbit [3], delivering appropriate pairs of images for processing. But since then, there are two Sentinel-1 satellites theoretically images from different satellites can be processed, because it would halve the time when satellites are in the same position, and could double the number of the suitable images.

After the image positions are selected, the polarization to be used should be decided. Usually the most reliable information on vertical deformation can be derived from the vertical-vertical (abbr. VV) polarization [10].

For this study two images were selected, a search for data has been performed for the 21 to 27 of August, 2016 period in the vicinity of the event, c.f. Table 2. There have been five images found (c.f. Table 3, Figure 4). Among them two images have been observed on ascending orbit (one 1A and one 1B), and the remaining three images on descending orbit (two 1A and one 1B).

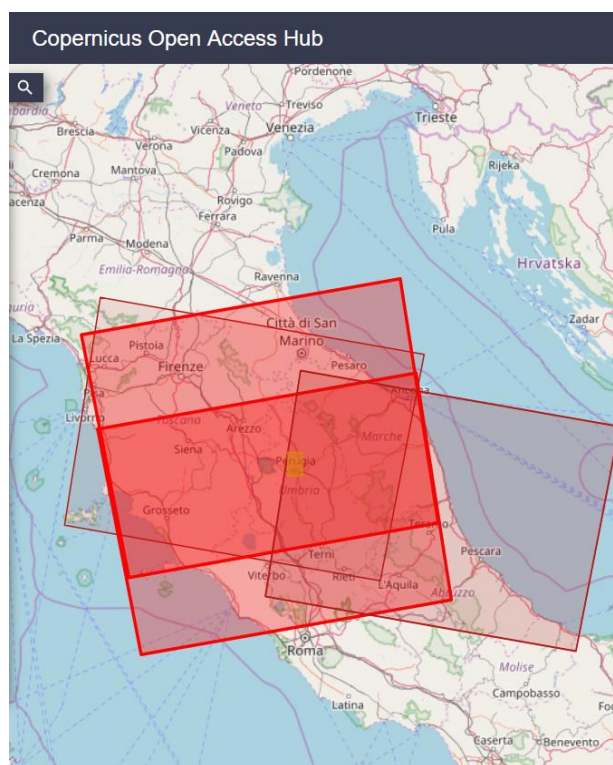


Figure 4. Search results and their positions

On figure 4 the two images, which were finally the source of the processing are highlighted. There were attempts processing the other three images. According to the tests it appeared that if the area of interest is close to the edge of the overlapping area, i.e. in the 10% border zone, then no vertical changes can be detected on the interferogram. Also, it turned out that images from descending and ascending orbits cannot be combined regardless the source of the image, i.e. identical or different satellites.

Two images a Sentinel-1A image on 2016-08-21 and a Sentinel-1B image on 2016-08-27 has been used:

- S1A_IW_SLC__1SDV_20160827T170542_20160827T170609_012789_014270_B9C6
- S1B_IW_SLC__1SDV_20160821T170509_20160821T170538_001718_002770_6482

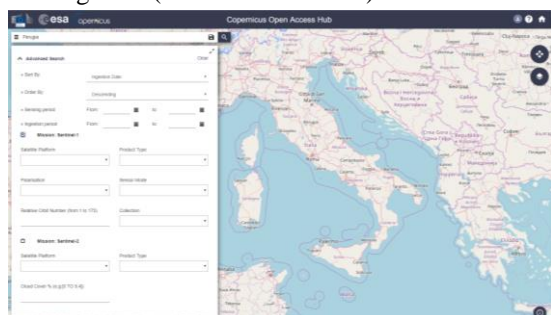


Figure 3. Scientific Data Hub

The main processing software was the Sentinel-1 Toolbox of the SNAP software, apart from the determination of the vertical deformation map from the interferogram, which can be performed in an independent step with the use of the Statistical-Cost, Network-Flow Algorithm for Phase Unwrapping (SNAPHU) program, every other step was performed by it.

The steps of the workflow can be seen below, c.f. Figure 5.



Figure 5. Processing workflow

During the first, *S1 TOPS Coregistration* step, the master and slave images has been chosen. Every image file (ref frame on Figure 6) contains three subsequent *Sub-swaths* (middle white frame on Figure 6), thus another selection has been made to choose that which part is the area should be processed. If necessary, the location can further be narrowed by using the tool *Bursts* (small white frames on Figure 6).

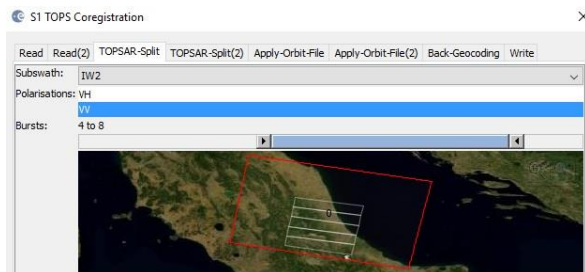


Figure 6. *S1 TOPS Coregistration* step with *Sub-swaths* and *Bursts*

The results are the intersecting parts of two datasets in the same position, as shown below c.f. Figure 7 and 8.

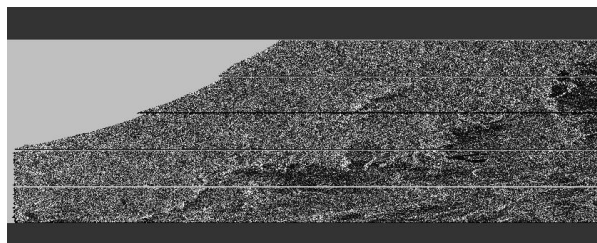


Figure 7. Bursts of master image

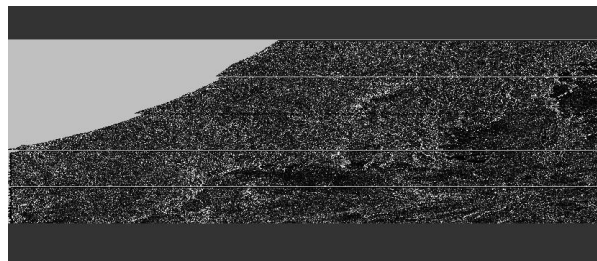


Figure 8. Bursts of slave image

In the next step, the *Interferogram formation*, the following corrections are applied [11]:

- $\Delta\phi_{flat}$: phase correction for the Earth curvature (often named as *flat Earth phase*),
- $\Delta\phi_{elevation}$: phase correction due to topography,
- $\Delta\phi_{displacement}$: surface deformation correction,
- $\Delta\phi_{atmosphere}$ phase correction accounting for atmospheric differences,
- $\Delta\phi_{noise}$: phase noise correction generated by temporal change of the scatterers, varying look angle, and volume scattering.

The corrections are simply summed to a correction term, $\Delta\varphi$

$$\Delta\varphi = \Delta\varphi_{flat} + \Delta\varphi_{displacement} + \Delta\varphi_{atmosphere} + \Delta\varphi_{noise}$$

where:

$$\Delta\varphi_{flat} = -\frac{4\pi}{\lambda} \frac{B_n s}{R \tan \theta}$$

$$\Delta\varphi_{elevation} = -\frac{\Delta q}{\sin \theta} \cdot \frac{B_n}{R_0} \cdot \frac{4\pi}{\lambda}$$

$$\Delta\varphi_{displacement} = +\frac{4\pi}{\lambda} d$$

In these equations B_n is the normal baseline, R_0 is the radar-target distance Δq is altitude difference, s is slant range displacement and θ is the radiation incidence angle with respect to the reference. (For the geometrical representation of some of these quantities, see Figure 9.)

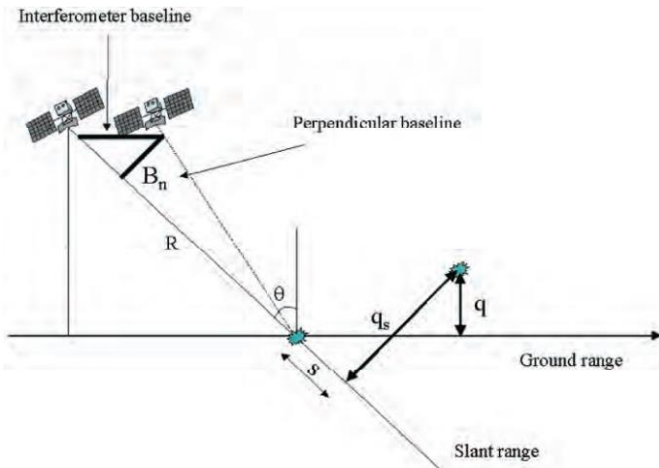


Figure 9. Sketch of the geometry of observing with a SAR satellite for the use of the InSAR technology. [12]

The *S1 Tops Deburst* function eliminates the horizontal lines on the images by merging the neighboring stripes.

To get an image where only the deformations remain, the topography effects must be subtracted by using the *Topographic Phase Removal* function (c.f. Figure 10.).

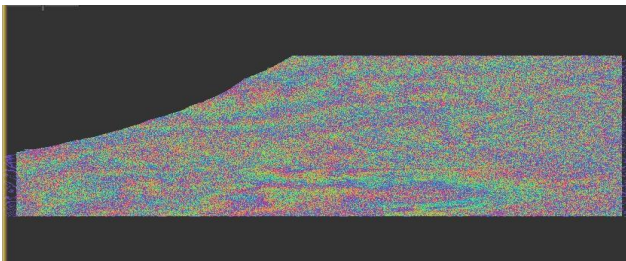


Figure 10. Phase image without topographic effects (red=0 blue=2π).

On the aforementioned figure (Figure 10.) the dotted nature is the consequence of the lack of coherence. Coherence loss can be occurred due to temporal and geometric decorrelations, volume scattering and processing errors. To reduce this noise, a filter should be applied. In the SNAP program this filter applies the Goldstein method, and the function utilized is the *Goldstein Phase Filtering*. The results of the filtering are shown on Figure 11.

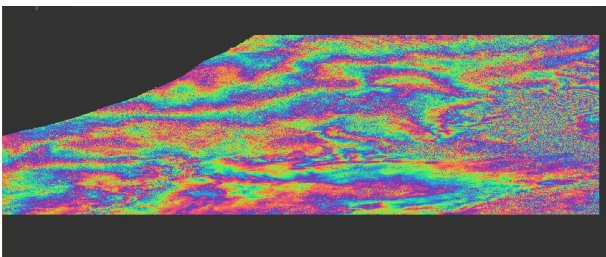


Figure 11. Phases after the Goldstein filter has been applied (red=0 blue=2π)

The improvements are conspicuous between the two processing states (Figure 10 vs. Figure 11.). The results are ambiguous since the values in every pixel are between 0 and 2π. In order to interpret the results for engineering purposes, the displacement values in meter dimension are preferred. In order to achieve this, a so-called *Phase Unwrapping* step should be applied. This is the step, which can be done in a separate software (SNAPHU) dedicated to this task. Phase unwrapping solves this ambiguity problem by integrating phase difference between pixels next to each other. (c.f. Figure 12.) [1]

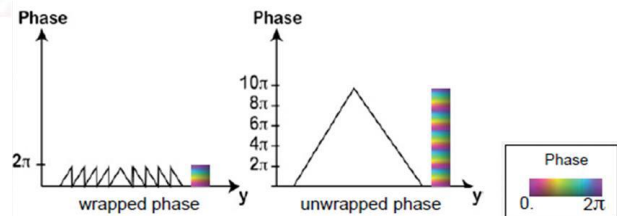


Figure 12. Phases before and after unwrapping. [1]

Subsequently, the *Phase to Displacement* function can be used for obtaining vertical displacement values as distances.

The two final steps, the *Update Geo Reference* and the *Ellipsoid Correction* is needed in order to transform the results into geographically correct coordinate system. (c.f. Figure 13.)

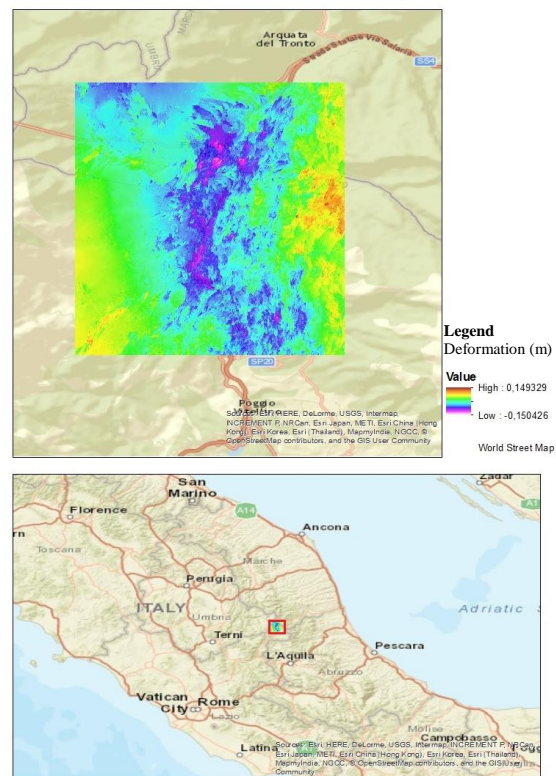


Figure 13. The displacement results in geographically correct position

IV. RESULTS

In the study an attempt has been done to combined use of Sentinel-1A and Sentinel-1B images. For the purpose a major earthquake has been analyzed as a case study. According to the investigations it was found that

- (1) If the area of interest is close to the edge of the overlapping area, i.e. in the 10% border zone, then no vertical changes can be detected on the interferogram.
- (2) Images from descending and ascending orbits cannot be combined regardless the source of the image, i.e. identical or different satellites.
- (3) Images of different satellites, 1A and 1B (with both being on ascending orbit) could efficiently be used for vertical deformation analysis.

The successful processing of 1A and 1B satellites ascending images has led to comparable results with other studies. According to our investigation, on the 24th of August, 2016, the area of Perugia due to the earthquake has been sunk about 14-15 cm. In a similar study published on Geo-Sentinel website [13], a subsidence of 20 cm has been determined. The difference arises from the different data they used, the different processing software, which involves several differences of processing, such as phase filtering method, or correction models.

The vertical changes follow the tectonic lines, in the lower areas subsidence, in the higher areas uplift can be detected, so the motions are consistent to the topography.

REFERENCES

- [1] Sentinel-1: ESA's Radar Observatory Mission for Copernicus Operational Services, *S1 Data Sheet*, homepage: http://esamultimedia.esa.int/docs/S1-Data_Sheet.pdf, retrieved 2016-10-18.
- [2] ESA, Copernicus: Observing the Earth, homepage: http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3, retrieved 2016-10-18.
- [3] L. Bányai, E. Szűcs, J. Kalmár, I. Eperné Pápai, D. Bán, "Az InSAR technológia alapjai és a reflektáló felületek jellemzői", *Geomatikai Közlemények*, vol. XVII, pp. 59-68, 2014.
- [4] M. Simons and P. A. Rosen, "Interferometric Synthetic Aperture Radar Geodesy", *Treatise on Geophysics*, vol. 3, pp. 391-446, 2007.
- [5] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, K.P. Papathanassiou, "A tutorial on synthetic aperture radar". *IEEE Geoscience and Remote Sensing Magazine*. vol. 1, pp. 6-43, 2013. doi:10.1109/MGRS.2013.2248301
- [6] R. F. Hanssen, "Satellite radar interferometry for deformation monitoring: a priori assessment of feasibility and accuracy", *International Journal of Applied Earth Observation and Geoinformation*, vol. 6, pp. 253-260, 2005.
- [7] <http://ds.iris.edu/ieeb/index.html?format=text&nodata=404&starttime=2016-08-24&endtime=2016-08-25&orderby=time-desc&limit=1000&maxlat=42.9790&minlat=42.4290&maxlon=13.7975&minlon=12.3555&zm=10&mt=ter>
- [8] D. Bihari, "Italian earthquake: The same can happen in Hungary" web article: <http://24.hu/kozelet/2016/08/24/olasz-foldrenges-magyarorszagon-is-megtortenhet-ugyanez/> retrieved 2016-10-18.
- [9] Sentinel-1 Toolbox of the Sentinel Application Platform (SNAP), available at the Science Toolbow Exploitation Platform (STEP); link: <http://step.esa.int/main/toolboxes/snap/>, retrieved 2016-10-18.
- [10] L. Bányai: personal discussion, 2016.
- [11] L. Veci, "Sentinel-1 Stripmap Interferometry: Sentinel-1 Toolbox Interferometry Tutorial", Array Systems Computing Inc., pp. 41, 2015.
- [12] K Fletcher (ed.), "InSAR Principles: Guidelines for SAR Interferometry Processing and Interpretation (TM-19, February 2007)", ESA Publications, ESTEC, Postbus 299, 2200 AG Noordwijk, The Netherlands, pp. 48, ISBN: 92-9092-233-8, ISSN: 1013-7076, 2007.
- [13] Geo-Sentinel <http://geo-sentinel.eu/meg-egyszer-az-olaszorszagi-foldrengesrol-sentinel-1-gyel-reszletesebben/>

Control of Mechatronic Devices using Computer Vision

Olga Baklanova*, Alexander Baklanov*, György Györök**, Mariya Yemelyanova*,
Zhenisgul Rakhmetullina* and Yevgeniy Baklanov***

* D. Serikbayev East Kazakhstan state technical university, Ust Kamenogorsk, Kazakhstan

** Obuda University, ARTF, Székesfehérvár, Hungary

Tomsk State University, Russia

O_E_Baklanova@mail.ru, ABaklanov@ektu.kz, gyorok.gyorgy@amk.uni-obuda.hu, e-maria@list.ru,
rahmetullina@mail.ru, miltenfiremage@gmail.com

Abstract—Modern CNC machines and robots usually work according to a certain program. External control can also be applied manually. However, during the execution of the task, there may be situations when adjusting the task is necessary. This work is devoted to the management and adjustment of mechatronic devices using computer vision. This paper describes the approach for quantifying loss of quality (in information value) of digital images when modifying their size.

I. INTRODUCTION

The industrial manufacturing of products from various materials is always accompanied by a certain degree of defect caused by various inadequacies in shape, hidden imperfections manifested during processing, as well as unfitness of work surfaces of the finished products to normal operation in the future.

Until recently, the implementation of quality control required, and in many enterprises still requires, the presence of controllers who carry out the evaluation visually. The process of sorting parts or products subject to control (with human participation) must be carried out directly on the production line in real time regardless of the production rate.

At present, a number of solutions for automated visual quality control in production are present in the market. Among the companies that represent solutions in this area are SICK [1, 2], Siemens [3], National Instruments [4], Microscan [5], Cognex [6], Sensopart [7], Barco [8] etc.

Foreign researchers such as T. Maenpaa [9], M. Pietikainen [10], A. Ahonen [11], T. Ojala [12], R. Haralick [13], H. Kauppinen [14], as well as researchers from CIS countries I.A. Kudinov [15], S.M. Sokolov [16], A.M. Bondarenko [17] all have made a big impact in this sphere.

The transition to automatic quality control is inevitable and requires not only the creation of special equipment, but also the development of appropriate mathematical and software systems for information processing. The application of said systems to automation of quality control can significantly improve the manufacturing efficiency.

Manufactured parts made by milling, grinding, as well as casting or stamping, may not be suitable for further use. This may be due to a mismatch of geometric characteristics, such as the contour, shape, dimensions of the part, or because the work surfaces do not meet the requirements for accuracy and presence of defects. The task of image

analysis is finding and recognizing all kinds of surfaces that are subject to control.

When analyzing such images, several types of problems arise, each of the problems requiring a definite solution. First, there are tasks associated with the preparatory stages. These include correct identification of areas of interest, the separation of surface of the part from the background (segmentation) [18], detection of textures and homogeneous areas. Secondly, these are problems relating directly to the analysis of surfaces, such as precise assessment of the surface type based on its texture and recognition of textures pertaining to different surfaces.

Real-time processing can be divided into two tasks: the main task of recording and storing images, and the task of processing images for information required for adjustments to the robot and to the production line.

In conditions of continuous recording and quality control of the production process based on image data, a memory overflow happens in the management system after some time. This paper proposes the use of new image compression algorithms to allow for the real-time processing of data from the production line, and thus also for the adjustments to the operation of the robot based on the appearance of new factors affecting the quality of production.

This paper describes the approach to quantifying the loss of quality (in information value) of digital images when modifying their size. Real-world practice of image analysis suggests that for most digital images a linear decrease in their size up to a certain threshold does not lead to the loss of required information. This is possible due to the uniform scaling of all informative elements of the image.

II. METHODS OF AUTOMATED ACQUISITION OF DIGITAL IMAGES

The automated acquisition of micrographs requires an automated microscope containing a motorized specimen stage, a mechanism to change the filters, a focusing mechanism and a revolver to change lenses. Typically, in these kinds of tasks the manufacturer's software is responsible for the control of the motorized units of the microscope. Simplified classification of microscopic cameras is shown in Figure 1. The preferable configuration is a trinocular microscope, with a camera that does not require an optical adapter and with a digital interface USB, controlled by a computer and supporting TWAIN.

The process of obtaining images can be carried out either without an operator, such as in the case of a fully automated microscope, or with the participation of an operator. In case of an automated microscope, the software fully controls the microscope and image analysis starts after recording the photo into the memory of a computer. It is a requirement for an analysis program that the results are collected to a database accessible over the LAN. From the database, the results will be read by a SCADA-system, and the system will make decisions based on it [19].

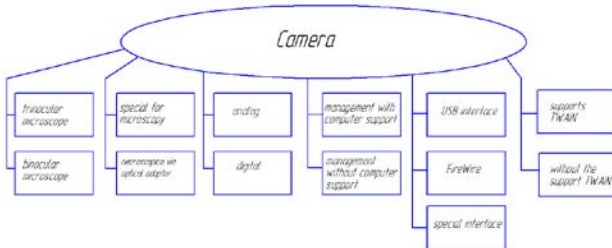


Figure 1. Classification of microscopic cameras

In case of a partially automated or non-automated microscope, the operator is responsible for changing the samples, the selection of the microscope objective, taking a photo to preserve on a workstation, running the image analysis. As in the previous case, the analysis program is required to enter the results into a database, where a SCADA-system can use it, and the system makes decisions [20].

III. METHODS OF REDUCTION OF IMAGES OF THE TECHNOLOGICAL PROCESS

In automated recognition systems, the following methods of reduction are commonly implemented [21]:

- None. There is no interpolation as such, the document sends the raw camera footage. Since there is no interpolation, you can expect high FPS rates: nothing slows it down.

- Nearest Neighbor specifies nearest-neighbor interpolation. The fastest method of interpolation that provides the image of the lowest quality compared to other interpolation algorithms.

- Bilinear specifies bilinear interpolation. No prefiltering is done. This mode is not suitable for shrinking an image below 50 percent of its original size.

- High Quality Bilinear specifies high-quality, bilinear interpolation. Prefiltering is performed to ensure high-quality reduction. A relatively fast interpolation method. The resulting image is really competitive.

- Bicubic specifies bicubic interpolation. No prefiltering is done. This mode is not suitable for shrinking an image below 25 percent of its original size.

- High Quality Bicubic specifies high-quality, bicubic interpolation. Prefiltering is performed to ensure high-quality reduction. This mode ensures the highest quality of the transformed images [22].

- WinScale algorithm for scaling images using the pixel model based on their area. The algorithm has low complexity: it uses no more than four pixels of the original image to calculate one pixel of the received image. The algorithm has a good performance – the output image has smooth edges, and the blur is variable [23].

To shrink an image, groups of pixels in the original image must be mapped to single pixels in the smaller image. The effectiveness of the algorithms that perform these mappings determines the quality of a scaled image. Algorithms that produce higher-quality scaled images tend to require more processing time. In the preceding list, Nearest Neighbor is the lowest-quality mode and High Quality Bicubic is the highest-quality mode [10]. WinScale algorithm results in the same quality as a bilinear algorithm in conditions of comparable complexity.

Based on the above, you would want to use the bicubic interpolation algorithm for scaling the image, considered the most optimal from the point of view of qualitative evaluation of the modified images and well-supported in all GPUs.

We propose two approaches to evaluating the reduction algorithms mentioned in this paper: comparative evaluation of the standard mean square error variance of the original and reduced images within a sliding window, and histogram evaluation [24].

IV. REDUCTION ALGORITHM BASED ON THE STANDARD MEAN SQUARE ERROR VARIANCE

The criterion is based on segmentation of the images into the same number of regions (disjoint windows), according to the selected step of segmentation, and calculating the root mean square error of brightness dispersion inside the window for the source and reduced images.

This algorithm consists of the following steps:

- 1) Original image is divided into square areas of equal size (the grid is superimposed on an original image);

- 2) Calculate the average intensity of every region (sum up intensity for every pixel and divide by the number of pixels);

- 3) For each region variance is calculated by the formula (1)

$$D_{i,j} = \frac{1}{m \cdot n} \sum_{x=1}^m \sum_{y=1}^n \quad (1)$$

where M, N are dimensions of the original image in pixels, m, n are dimensions of the selected windows in pixels, $L_{x,y}$ is the brightness of a pixel with coordinates (x,y) , $M_{i,j}$ is the average brightness within the window with coordinates (i,j) ;

- 4) Reduce the original image, and repeat steps 2 – 3. The scale and the variance value are recorded at each iteration of the cyclic reduction. Data is accumulated into arrays.

- 5) Calculate the loss of the information content according to the formula (2)

$$\Delta D = \frac{1}{I \cdot J} \sum_{i=1}^I \sum_{j=1}^J \quad (2)$$

where $D_{i,j}$ is the variance inside the window of an original image with coordinates (i,j) , $D'_{i,j}$ is the variance inside the window of a modified image with coordinates (i,j) , I is the number of rows of windows, J is the number of columns of windows. The data is stored in an array.

- 6) Loss of quality from the zoom is plotted based on the value of ΔD . The source image is considered to be 100% quality. Example of the algorithm is shown on Figure 2.

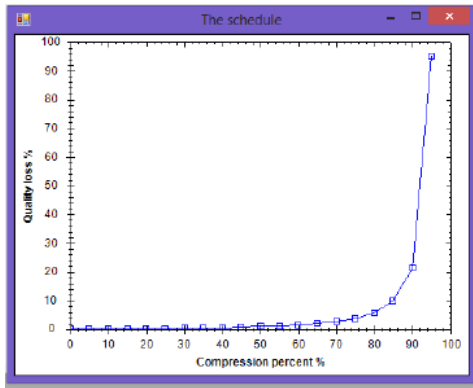


Figure 2. - The dependence of image quality loss for the algorithm based on average-squared error estimation variance

V. REDUCTION ALGORITHM BASED ON THE HISTOGRAM EVALUATION

The method is based on a comparison of "shapes" of brightness histograms for the original image and the scaled image. The standard error deviation of a source image histogram from the modified image histogram serves as a criterion.

Histograms for each component of the color space allow us to estimate characteristics of the digital image in terms of form of the color-brightness settings distribution. The cardinality of a histogram array is the same for any image [25]. This can be used to assess the differences between the original image and the scaled one. The standard error deviation of a source image histogram from the modified image histogram serves as a criterion. Histograms are constructed in relative frequencies, which leads to the difference in the total number of pixels of the original image and the scaled image:

$$\Delta G = \sqrt{\frac{1}{N} \sum_{i=1}^N (HF_i - HM_i)^2}, \quad (1)$$

where N is an array dimension histogram, HF is an array of relative frequencies of the brightness histograms for the original image, HM is a similar array for the modified image. Figure 3 shows the dependence of the image quality loss on the compression.

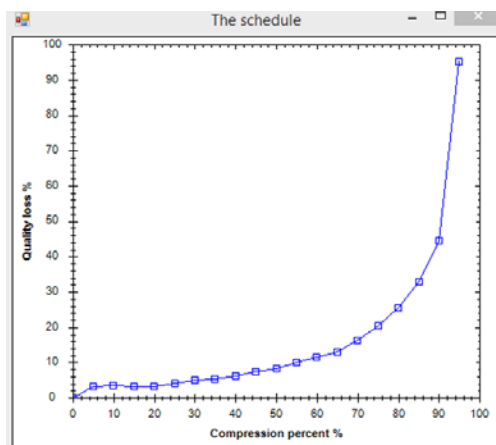


Figure 3. - The dependence of image quality loss for the algorithm based on the histogram evaluation

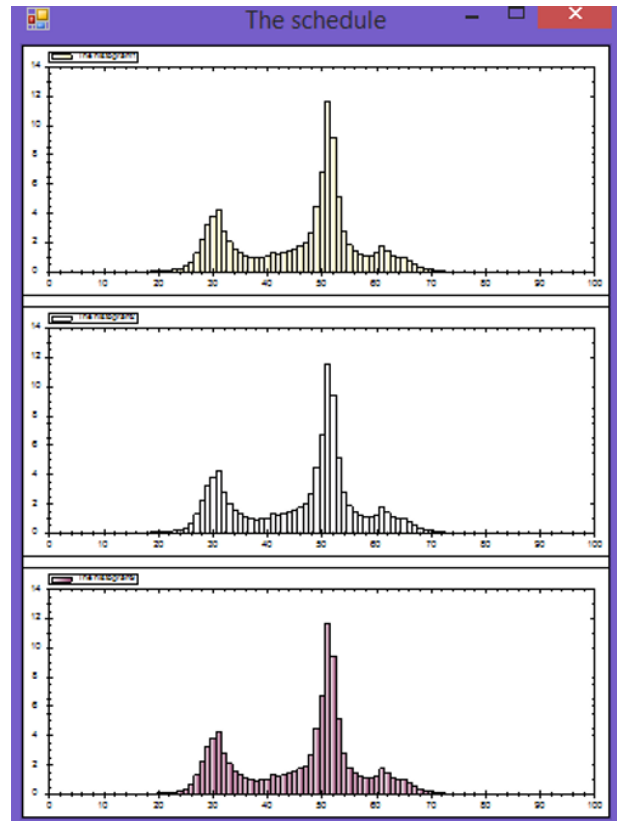


Figure 4. - The algorithm is based on the histogram evaluation

A software module was developed to analyze the proposed evaluation methods. Different types of mineral rock images were analyzed for changes in the average square error of the original and the reduced image. It was confirmed The viability of the proposed information loss estimation method has been proven for all analyzed images.

VI. RESULTS AND DISCUSSION

Joint analysis of the visual changes towards "information value" decrease for the given image leads to an increase in criterion value. This fact can be used to automatically calculate the reduction threshold for the original image. In the process the fixed values of the criteria are translated into a percentage ratio of information loss.

To do this the original image is modified in a loop by reducing its size by step q (q = 5%), and at each iteration a fixed criterion value is recorded. The cycle stops when the image scale is reduced to 0, i.e. the image is "degenerate", fitting 100% information loss. Based on this, the percentage ratio of information loss is calculated for each recorded step of the reduction. After that, the allowable percentage of quality loss, and thus the allowed threshold, are selected according to a chosen analysis methodology.

Testing was performed on the FESTO training and production robotized line. Mechatronics FESTO is the synergistic combination of mechanical engineering, electrical engineering, electronics, information technology and system analysis utilized in the design of products and automation processes (Figure 5).



Figure 5. - Robotic FESTO line

VII. CONCLUSION

The acquired information can be used in practice to increase the speed of recognition algorithms for digital images of the conveyor products. Reducing the size of the image N^2 times leads to an increase in processing speed of N times, which significantly increases efficiency by allowing the use of more "expensive" in terms of time, but more quality algorithms.

A software module was developed to analyze the proposed evaluation methods. Different types of mineral rock images were analyzed for changes in the average square error of the original and the reduced image. It was confirmed The viability of the proposed information loss estimation method has been proven for all analyzed images.

ACKNOWLEDGEMENT

The study has been conducted with the financial support of the Science Committee of RK MES in the framework of the program target financing for the 2017-2019 biennium by the program 0006/PTF-17 "Production of titanium products for further use in medicine".

REFERENCES

- [1] Automotive industry and production of components: the site is SICK. [Electronic resource]. URL: <https://www.sick.com/ru/ru> (accessed 12.09.2017)
- [2] O. N. Lysenko Machine vision from SICK/IVP - automation in industry, No. 3, 2007. pp. 30-33.
- [3] Automation systems for all requirements: website Siemens. [Electronic resource]. URL: <https://www.siemens.com/global/en/home/products/automation/systems.html> (accessed 12.09.2017)
- [4] Automated testing in the automotive industry: the website of the company National Instruments. [Electronic resource]. URL: <http://www.ni.com/ru-ru/innovations/automotive/automated-automotive-test.html> (accessed 12.09.2017)
- [5] Machine vision systems: the website of the company Microscan. [Electronic resource]. URL: <http://www.microscan.com/en-us/products/machine-vision-systems> (accessed 12.09.2017)
- [6] Machine vision: Cognex website. [Electronic resource]. URL: <http://www.cognex.com/products/machine-vision/?pageid=14404&langtype=1049> (accessed 12.09.2017)
- [7] Industrie 4.0: the website of the company Sensopart. [Electronic resource]. URL: <http://www.sensopart.com/en/industry-4-0> (accessed 12.09.2017)
- [8] Automotive industry: the website of the company Barco. [Electronic resource]. URL: <https://www.barco.com/ru/markets/Automotive> (accessed 12.09.2017)
- [9] Mäenpää T, Pietikäinen M. Classification with color and texture: jointly or separately? *Pattern recognition*, 2004, 37 (8), 1629-1640.
- [10] J Chen, S Shan, C He, G Zhao, M Pietikainen, X Chen, W Gao. WLD: A robust local image descriptor. *IEEE transactions on pattern analysis and machine intelligence*, 2010, 32 (9), 1705-1720.
- [11] Ahonen T, Matas J, He C, M Pietikäinen. Rotation invariant image description with local binary pattern histogram fourier features. In: *Image analysis*, 2009, 61-70.
- [12] T Ojala, M Pietikäinen, T Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In: *European Conference on Computer Vision*, 2000, 404-420.
- [13] R. M. Haralick. Propagating covariance in computer vision. *Performance Characterization in Computer Vision*, 2000, 95-114.
- [14] Silven, O., Niskanen, M., Kauppinen, H.: Wood inspection with non-supervised clustering. In: *Machine Vision and Applications*, 2003, 13(5-6), 275-285.
- [15] Kudinov I. A. and others Algorithm for panoramic image generation from multiple cameras with overlapping fields of view and its implementation: proc. In: *7-th scientific-technical Conf. "Technical vision in control systems-2016" (TSSU-2016): collection of abstracts*. Moscow, IKI ran, 15-17 March 2016. M.: IKI, 2016. P. 56.
- [16] Sokolov S. M. STZ in the control loop of the CNC machine. *The Sixth all-Union conference on control in mechanical systems*. Abstracts. – Lviv: Lviv scientific library of the Ukrainian Academy of Sciences, 1987. – P. 144.
- [17] Bondarenko M. A., Drinkin V. N. Assessment of informativeness of the combined images in multispectral machine vision systems. In: *Software systems and computational methods*. 2016. No. 1. P. 64-79. DOI: 10.7256/2305-6061.2016.1.18047.
- [18] Baklanova, O.E., Shvets, O. Ya.: Methods and Algorithms of Cluster Analysis in the Mining Industry. Solution of Tasks for Mineral Rocks Recognition. In: *SIGMAP – 2014, Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications*. SCITERPRESS. DOI: 10.5220/0005022901650171 (2014)
- [19] Baklanova, O.E., Shvets, O. Ya., Uzdenbaev, Zh.Sh.: Automation System Development For Micrograph Recognition For Mineral Ore Composition Evaluation In Mining Industry. In: *AIAl - 2014, Artificial Intelligence Applications and Innovations*, IFIP Advances in Information and Communication Technology, vol.436, 2014, pp. 604 -613. SPRINGER. DOI 10.1007/978-3-662-44654-6
- [20] Baklanova O.E., Baklanov, A.E., Shvets O.Ya: Methods and algorithms of computer vision for automated processing of mineral rocks images In: *Applied Computational Intelligence and Informatics (SACI-2015)*, 2015 IEEE 10th Jubilee International Symposium on 21-23 May, 2015, pp. 449 – 454, DOI:10.1109/SACI.2015.7208246
- [21] How to: Use Interpolation Mode to Control Image Quality During Scaling. [http://msdn.microsoft.com/ru-ru/library/k0fsyd4e\(v=vs.110\).aspx](http://msdn.microsoft.com/ru-ru/library/k0fsyd4e(v=vs.110).aspx)
- [22] Interpolation Mode Enumeration. [http://msdn.microsoft.com/ru-ru/library/system.drawing.drawing2d.interpolationmode\(v=vs.110\).aspx](http://msdn.microsoft.com/ru-ru/library/system.drawing.drawing2d.interpolationmode(v=vs.110).aspx)
- [23] Chun-Ho Kim, Si-Mun Seong, Jin-Aeon Lee, and Lee-Sup Kim: Winscale: An Image-Scaling Algorithm Using an Area Pixel Model In: *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, VOL. 13, NO. 6, JUNE 2003 549
- [24] Baklanova, O.E., Shvets, O. Ya.: Development of Methods and Algorithms of Reduction for Image Recognition to Assess the Quality of the Mineral Species in the Mining Industry. In: *ICCVG-2014, Computer Vision and Graphics, Lecture Notes in Computer Science*, vol. 8671, 2014, pp. 75-83. SPRINGER. DOI 10.1007/978-3-319-11331-9.
- [25] Gonsales, R. C., Woods, R. E.: *Digital image processing*, 3rd edition, Pearson Education, 2011, 976 p.

An algorithm for object classification procedure for ISAR images

K.O. Slavyanov*, C.N. Minchev**

* National Military University/Department “Computer Systems and Technology”, Shumen, Bulgaria

** University of Shumen/ Faculty of Technical Sciences, Shumen, Bulgaria

k.o.slavyanov@gmail.com

chavdar_minchev@yahoo.com

Abstract— This article offers a neural network architecture for automatic classification of Inverse Synthetic Aperture Radar objects represented in images with high level of post-receive optimization. A full explanation of the procedures of two-layer neural network architecture creating and training is described. The classification in the recognition stage is proposed, based on comparison with flying objects images from a database. The classification sets are gained by distinctive specifications in the structural models of the aircrafts. The neural network is experimentally simulated in MATLAB environment.

I. INTRODUCTION

For the classification systems of Inverse Synthetic Aperture Radars (ISAR) the neural networks technologies for better image reconstruction are proven to be successful [1]. An opportunity for improved information analysis in that area is suggested to be the development of better algorithms for recognition of the various flying objects.

In [2] to solve the problem of recognition of not cooperating objects of observation, after acquiring radar image, an algorithm based on fuzzy logic can be used to make this classification with a high degree of credibility while controlling the error rate. The effect of uncertainty in the identification process is reduced if it can be trained or if the experience of expert can be studied [7]. Many opportunities are revealed for in-depth research and implementation of new ideas and approaches to accelerate the process of implementing the principles of inverse aperture synthesis in practice. At this stage a common standard for assessing the quality of the radar image is not created [5]. New methods for detecting and analyzing specific characteristics of objects in ISAR - images of moving objects [4] are needed in order to differentiate them into different classes. An algorithm for recognition of objects in the radar image by comparison with standard contour models of planes is presented in this paper.

II. PRECONDITIONS

For the simulation environment is assumed that the process of obtaining a horizontal orientation of the observed object in a network of 256x256 pixels is completed with linear resolution at azimuth and distance, respectively $\Delta L = 0.5 [m]$ and $\Delta R = 0.5 [m]$. Procedures for filtering the resulting image and extraction of 128x128 pixels subarea containing the object silhouette and image optimization are also preconditions for the binary matrix S with the aircraft object [6,7,8].

Contour object patterns are placed exactly in the middle of the frame, both horizontally and vertically. Model matrices with the size of 128x128 elements are formed as follows: If the pixel of the graphical contour model is part of the contour model, the value of 1 is assigned to the corresponding matrix element, otherwise the element is 0.

For the simulation experiment, sixteen airplane models are defined in a rectangular network of 128x128 pixels with network dimensions. Sixteen exemplary graphical contour models are used: Eurofighter Typhoon, Pilatus 9M, Rafale, Mirage 2000, MiG-29, Gripen, Falcon 2000, F-22, F-18, F-16, C-130 H, Bombardier Q400, Boeing-747, Boeing-737, Boeing-707 and Embraer Legacy 600. The reference models are created on detailed graphical maps, accompanied by precise data on the geometric dimensions of objects in the three dimensions. Graphics cards and data are published on the FAS website (Federation of American Scientists). The models are designed to be proportional 2D schemes of real aircrafts (fig.1).

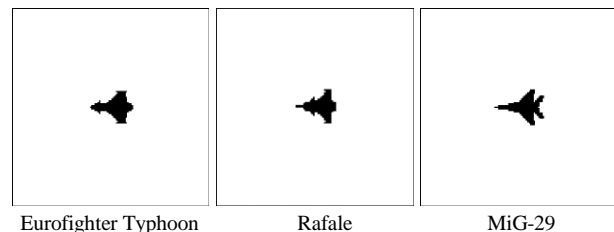


Figure1. Reference models from aircrafts database.

The modeling is carried out under the following initial conditions. Objects are observed with ISAR, their movement is simulated in a rectilinear trajectory at constant speed and at constant altitude during observation. The object is modeled in its own two-dimensional coordinate system with network dimensions on both coordinates [3].

It is assumed that high resolution at a distance is realized in the ISAR by usage of impulses with linear frequency modulation. ISAR pulses are designed with linear frequency modulation of duration $T = 10^{-6} [s]$, repetition period $T_p = 10^{-5} [s]$ high frequency oscillation $f = 10^{10} [GHz]$, wavelength $\lambda = 0.03 [m]$, and a full frequency deviation $2\Delta F = 3.10^8 [Hz]$ [6].

It is assumed that the positions of some of the scatterers of the three-dimensional variations of the reference

patterns are located with shading effect generation on other scatterers arranged behind them in the course of irradiation with the emitted radar signal.

A model of reconstructed ISAR image of the flying object in 128x128 pixel grid is used for the experiment in presence of Gaussian white noise with constant zero mean and variance 0.01 and "salt and pepper" noise with density 0.015. The additive "white" noise and impulse interferences produced by the peak noise are presented on figure 2. The experiments are developed in MATLAB environment.

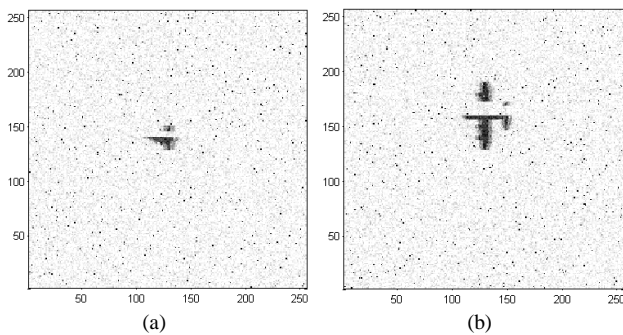


Figure 2. Reconstructed images in presence of additive noise for the aircrafts Rafale (a) and C-130 H (b).

It is presumed that the ISAR image is processed with automatic focusing procedure locked on the final image [4].

With this algorithm, the object is accurately detected in the radar image by comparison with contour reference aircraft models from the database.

Numerous methods for extracting characteristic lines (contours) and contours from the structure of an image, which are based on the evaluation of the first and second derivative of the function of the intensity, are known in the theory of digital image processing [8].

Based on the fact that the images that are being processed at this processing stage are in binary format when there is only one monolith object within the frame, a filtering procedure for drawing the contour is applied to the images from the database (Fig.3).

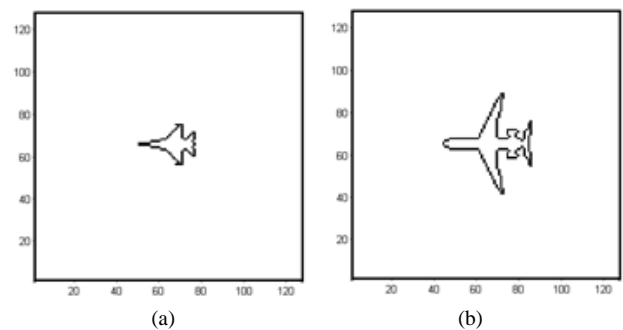


Figure 3. Extracting the contour line of the object through a filtering procedure in the images of F16 (a) and Embraer Legacy 600 (b).

In the filtration of the resulting ISAR image, various image clearing procedures are used to form the structure of the filtration system and digital image processing [9]. Their purpose is to provide an image extraction and a contour image of optimal quality that is adequately

corresponded to the shape of the subject in the frame, even in case of highly noisy images. Solving this task is of particular importance for the next step of processing the information associated with the recognition of the final object.

III. AN ALGORITHM FOR OBJECT RECOGNITION PROCEDURE

The availability of a ready database with a large amount of detailed models of flying objects (their characteristics, features of their structure) is essential for rapid decision on the specific task.

A neural network of type "Backpropagation" is chosen for the algorithm described before.

The problem is solved in converting the input image into a vector that can be classified by the neural network, similar to one of the classes (models) in a database, formed previously. A number of 16 etalon models, with dimensions of 128x128 pixels, is chosen for the comparison. Patterns are represented by binary matrices whose elements are numerical expression of the graphic-described solid models of aircraft with a known geometry. It is considered in this procedure that the possible error is within 2 pixels in eight directions. To remove the ambiguity of the subject in position, twenty-five supporting matrices are formed for each model by translation of the etalon model at distance of 2 pixels in eight directions on the center of the image (fig. 4). The etalon models database is formed by these 400 matrices (16x25). Figure 4 illustrates the principle of creating matrices $W_{M,1}, W_{M,2}, \dots, W_{M,N}$ for a reference model number corresponding to a Eurofighter Typhoon.

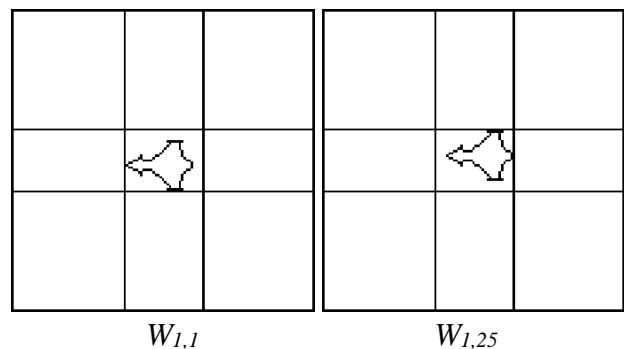


Figure 4. Graphic representation of matrix elements $W_{M,N}$ for model with number $M = 1$ (Eurofighter Typhoon).

In the next stage of the chosen designing approach, the etalon matrices with the values of pixel intensity are reshaped in vertices of 128x128=16384 element so each matrix is transformed to one column. These vertices are formed in one matrix of "training", called Training (16384x400). 400 is the number of objects in the database, multiplied by twenty-five items, which are subject to the procedure for recognition, a 16,384 is the number of pixels in an image. At this stage, a matrix for the "desired result" named "Target" is also constructed, which is necessary for the neural network process of training. The matrix has a dimension 16x400 - 16 rows of available sites classified by their solid silhouettes and 400 columns, because each etalon model is represented by twenty-five of his positions. The location of the non-zero element of

each column corresponds to the number of class (line) with which the result of recognition is associated.

In accordance with the algorithm proposed before, a

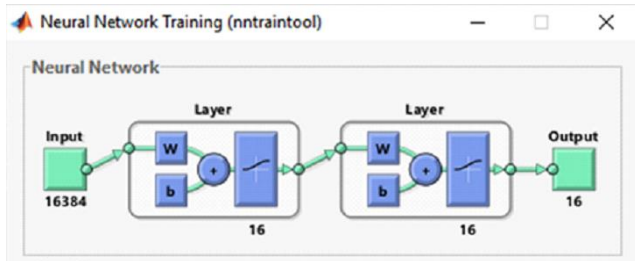


Figure 5. Block diagram of a neural network, designed in MATLAB environment.

neuron architecture consisting of two layers is designed by means of the Matlab programming language (fig.5) and is modeled in Simulink environment (fig.6). In theoretical aspect, the learning process of this type of neural networks has proven convergence and therefore in a sufficiently long period of self-training, the neuron's weights should be suitably adjusted to produce correct classification of the vectors from the training sample. The first layer of the neural network is "hidden" and is composed of 16 neurons with a log-sigmoid transfer function. These neurons form subclasses, some of which the input vector is classified with. The internal structure of this layer is depicted on figure 6.

A line called "Delays 1" that converts the elements of the input sequence into an input vector is in the layer structure. Log-sigmoidal transfer function provides a high

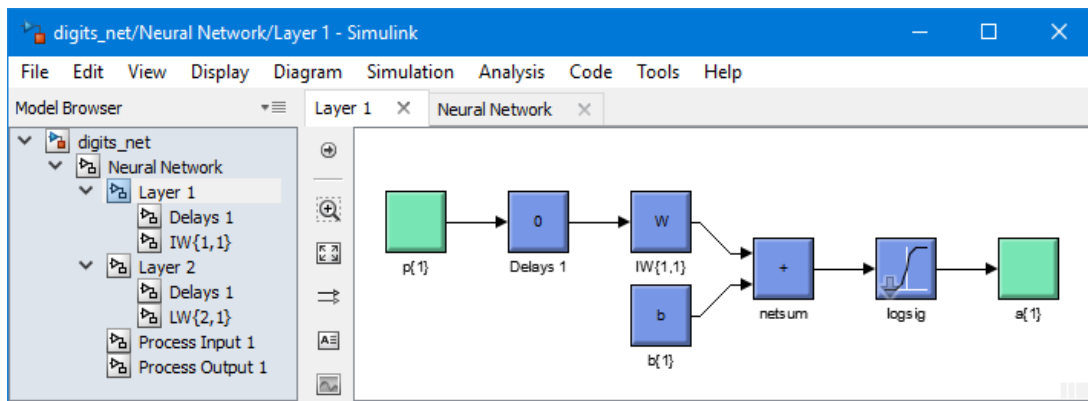


Figure 6. Structure of the first layer of the neural network built in Simulink.

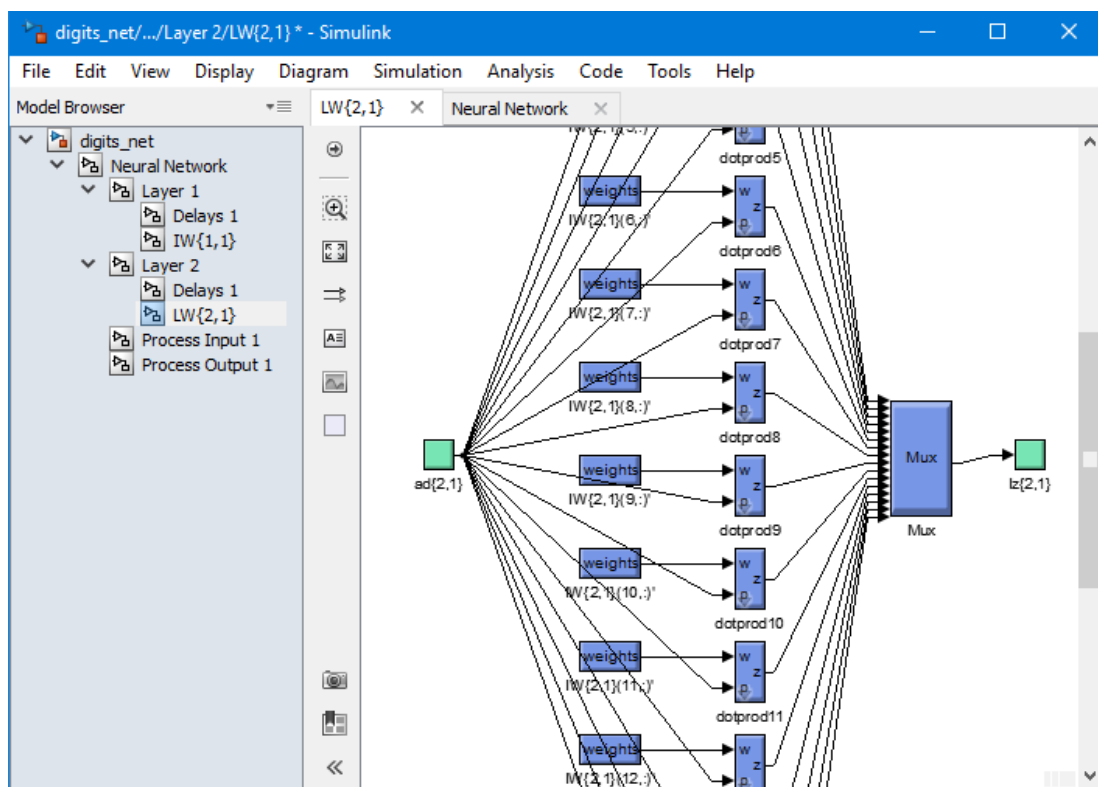


Figure 7. Structure of the weight matrix at the entrance of the second layer of the neural network.

sensitivity and high resolution in the recognition process.

The weight matrix IW consists of sixteen weight vectors called "weights", whose specific values are determined in the education stage of neural network. The second layer according to the final number of desired classes is designed to have 16 neurons. The number of the "winner" neuron is corresponding to that one of the all solid aircraft models to which the current input vector is brought to. The role of this layer is to process a classification of the first layer results and to generalize them to a fixed number of user classes (16). Its structure is similar to the structure of the first layer. The expanded structure of the inputs of the neurons in the layer is presented on figure 7, wherein the weight matrix LW is composed of sixteen weight vector, whose specific weights are determined in the training stage of the neural network.

A training of the neural network is processed for the next stage of the neural architecture realization, which is essentially an adjustment of coefficients of the weighting matrices of neurons of the two layers. Embedded algorithms and procedures are used for automated self-training of the Matlab neural networks. A method for training a neural network with "teacher" in accordance with the following sequence of actions is processed. The initial training of the neural network is carried out free of interference.

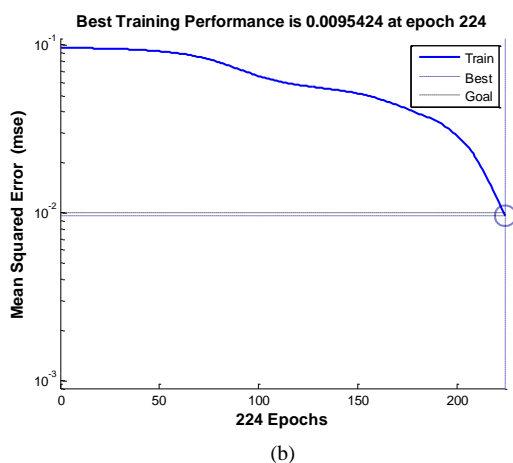
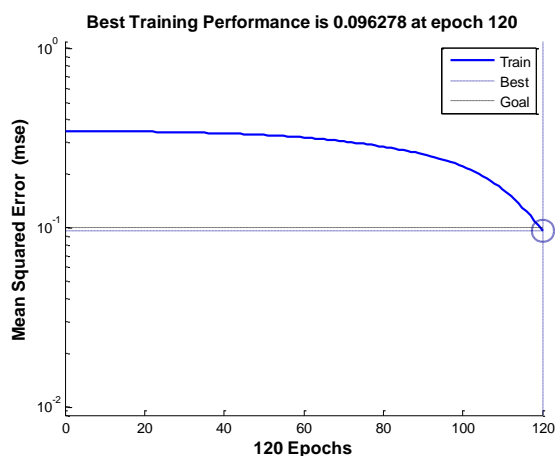


Figure 8. Desired possible error 0.1 during the training in noise free environment is reached at 120 epochs – (a). Desired possible error 0.01 reached at 224 epochs – (b).

The input of the network is fed with "training" matrix Training1, containing the values of the intensities of the pixels obtained from the reference models. The desired result indicated by the matrix Target is the product of the output of the network is designed to produce. Back propagation of error is the learning algorithm used. The goal for the possible error is chosen to be 0.1 and the error calculating function is of the type sse (sum squared error - accumulated value of the square error). The maximum of the training epochs is limited to 1000. The training results are illustrated on Figure 8 (a).

For the next step in the network training process higher requirements are implemented – the goal for the possible error is chosen to be ten times lower now - 0.01. The results of that training are shown on Figure 8 (b).

In line with the graphics on figure 8 the desired threshold is reached in two cycles of training in which the learning process is considered to be complete. Modeling the process of synthesis of this neural architecture is carried out in Matlab environment. The results of the neural network for the ISAR observed object "Falcon 2000" (fig.9) are presented on figure 10 where the position of the etalon model for that airplane is 7 and the object is properly classified.

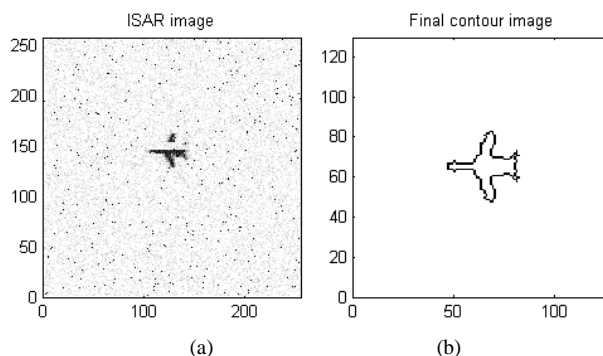


Fig. 9. ISAR image received (a) and optimized (b).

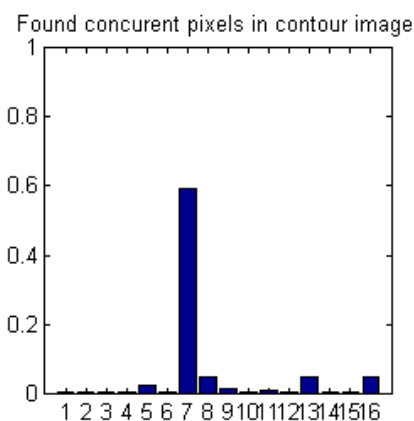


Figure 10. According to the neural network classification the object is recognized as the airplane Falcon 2000.

IV. CONCLUSION

In this article neural networks algorithm for object recognition procedure in ISAR image is designed. As a result of the analysis and the carried out experiments the following conclusions can be made:

The chosen decision making algorithm is logical and accurate for the class belonging of the observed object.

The described neural network operates like an associative memory and makes correct classification of the ISAR objects in high level of noise environment as well as if the objects are not full or heavily damaged.

The used number of neurons in the first layer is smaller than in other networks because it depends of the chosen models in contrast to the image pixel number.

REFERENCES

- [1] F. Benedetto, F. Riganti Fulginei, A. Laudani, G. Albanese, *Automatic Aircraft Target Recognition by ISAR Image Processing based on Neural Classifier*, International Journal of Advanced Computer Science and Applications, Vol. 3, No.8, pp 98-103, 2012.
- [2] Boulay T., Lagoutte J., Mohammad-Djafari A., Gac N., *A Fuzzy-Logic Based Non Cooperative Target Recognition*, Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference, Nov. 2012.
- [3] Lazarov A.D., Minchev C.N., *Correlation-autofocusing-spectral 2-D ISAR image reconstruction from linear frequency modulated signals*, Digital Avionics Systems Conference, 2002. Proceedings, pp. 27-31, Oct. 2002.
- [4] Lazarov A. D., *ISAR Signal Modeling and Image Reconstruction with Entropy Minimization Autofocussing*, 25th Digital Avionics Systems Conference, IEEE/AIAA, 2006.
- [5] Li, J., Wu R., Victor, C.C. *Robust autofocus algorithm for ISAR imaging of moving targets*. IEEE Trans., AES, vol.37, No3, July 2001.
- [6] Minchev C. N., Slavyanov K.S., *An opportunity for improved modeling and information analysis in ISAR systems*, "Machines, Technologies, Materials", ISSN 1313-0226. ISSUE 7/2013, pp. 24-28, 2013.
- [7] Tan Y., Yang J., Li L., Xiong J., *Data fusion of radar and IFF for aircraft identification*, Journal of Systems Engineering and Electronics (Volume: 23, Issue: 5, Oct. 2012), pp. 715 – 722, 2012.
- [8] Image Processing Toolbox User's Guide, Mathworks, 2015.
- [9] Minchev, C.N., Slavyanov, K.O., *An algorithm for ISAR image optimization procedure*, Bulgaria, National Military University, International scientific conference, 2015.

A Closely Located Dual-Band FSS Design for X Band Applications

S. Ünalı*, H. Bodur**, S. Çimen** and G. Çakır**

* Bilecik Şeyh Edebali University/Department of Electrical and Electronics Engineering, Bilecik, Turkey

** Kocaeli University/Department of Electronics and Communications Engineering, Kocaeli, Turkey
sibel.unaldi@bilecik.edu.tr, hande.bodur@kocaeli.edu.tr, sibeltunduz@kocaeli.edu.tr, gonca@kocaeli.edu.tr

Abstract—This study focuses on single layer dual band Frequency Selective Surface (FSS) with closely located resonant frequencies for X band applications. The designed FSS comprises of cross dipoles and fragments of broken rings printed on a single layer dielectric material to get dual band frequency response. The proposed structure offers the advantage of distinguishing two closely located frequencies such as 8.8 GHz and 11.2 GHz. The frequency ratio of the upper resonant frequency to the lower one is as small as 1.27s. Furthermore, the thickness of proposed FSS is only $0.017*\lambda$ where λ is the wavelength of the lower resonant frequency. In addition to these properties a stable frequency response is observed for different incidence.

I. INTRODUCTION

Frequency Selective Surfaces (FSSs) are periodic structures which contain slots or metallic patches with different geometry to provide filter characteristics [1], [2]. FSSs have been extensively studied and used in various applications such as spatial filters, absorbers, polarizers, radomes, antenna reflectors, EBGs (Electromagnetic Band Gaps), AMCs (Artificial Magnetic Conductors), electromagnetic shielding, RCS reduction controlling practice [2]-[4]. Recently, with the rapid development of electronics and communication technologies increasing demands have led to research [2]. Multiband or dual-band FSS can be practice where multiple transmission bands are required such as multi-channel systems [5],[6].

One of the considerable property of FSS design is the sensitivity to the different incidence angle of the EM wave. In [7] and [8] angular stable performance is obtained due to the symmetric geometry of the unit cell design. Also in [1] closely spaced planar dual-band FSS is studied with the feature of the 90° rotation technique for the unit cell makes available to independence of incident polarization and angle. Several methods have been investigated to achieve closely located frequency characteristics. The closely band response was achieved by reducing the mutual coupling with the redistributed maximum current densities along the unit cell in [6]. In [9] and [10] closely spaced operating bands are provided with the meander lines printed on a single layer dielectric material. In some multiband FSS studies, double square loop [11] and concentric ring [12] elements investigated to obtain closely spaced operating bands. However, well-adjusted gaps between those elements produces closely spaced bands and very small gaps have difficulties in implementation.

In this paper, we present a closely located dual-band FSS which is constructed as an ultra-thin structure. The proposed structure has the advantage of a low frequency ratio of 1.27s. Also the design shows the stable performance to the various incidence angles.

In section 2 the geometry of the proposed single layer, two surfaces FSS and design parameters have been studied in detail. In section 3 the simulation results of the transmittance and reflectance are given for both Transverse Electric (TE) and Transverse Magnetic (TM) polarizations. Also surface current distributions are given at two stop bands for both TE and TM polarizations to show resonating metallic arms of the unit cell, and finally section 4 shows the concluding views.

II. DESIGN OF FSS STRUCTURE AND PERFORMANCE

This section presented FSS unit cell with simple geometry is demonstrated in Fig. 2. The metallic lines, printed on the dielectric substrate are consist of two separated parts on two surfaces. Part one is composed from a small cross dipole in the center with the fragments of a broken ring on the front surface. The second part on the back surface is derived by scaling the first part by k and rotating it by 40 degrees but the fragments of a broken ring are rotating -5 degrees again. The preferred angle of rotation provides the geometry that causes the parts to be separated from each other and also contributes to achieving angular stable frequency performance. While outer diameter of the fragments of a broken big ring is named R , smaller one is r . The lengths of the big and small cross dipoles are R and r , respectively. When part two is obtained, k is used as a scale factor between R and r . The frequency ratio can be adjusted with different values of k .

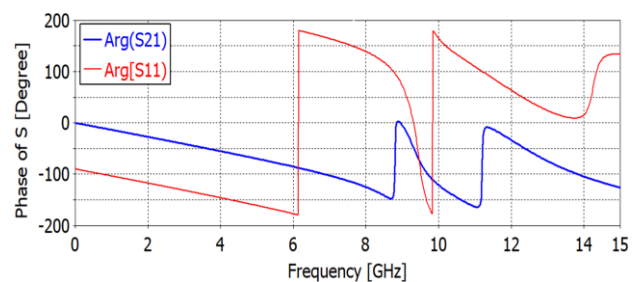


Figure 1. Simulation results of transmission and reflection phase at normal incidence

In Fig. 1, transmission and reflection coefficient phase characteristics are shown. The presented FSS has two operational frequencies ($f_1=8.8$ GHz and $f_2=11.2$ GHz) which is understood from Phase of transmittance curve obviously.

TABLE I.
DESIGN PARAMETERS OF THE UNIT CELL STRUCTURE

Design parameters	Dimensions (mm)
D	7
R	$k*r$
r	4.95
g1	0.22
g2	$k*g1$
h	0.508
k	13/11
m	1.65786
n	1.2899

And the design parameters of the unit cell which are shown in Fig. 2 are listed in Table 1. The unit cell designed on an Arlon AR 600 dielectric substrate with a thickness of (h) 0.508 mm, and a relative permittivity of (ϵ_r) 6. The total lengths of the unit cell are 7 x 7 mm.

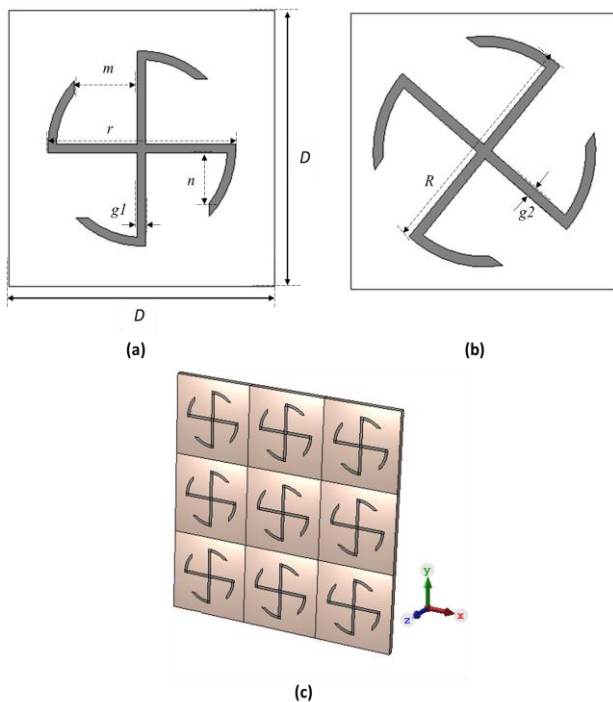


Figure 2. (a) Top view of the proposed unit cell, (b) Bottom view of the proposed unit cell, (c) Designed FSS structure

The gray region shown in Fig.2 represents the printed metallic lines and the white region represents the dielectric substrate. g1 and g2 are the width of the metallic lines

which illustrated in Fig.2 (a) and (b). In Fig.2 (c) perspective view of designed FSS structure is given.

Part one causes the higher operating frequency (f_2) and part two causes the lower operating frequency (f_1). Thus, the band spacing can be easily reduced or increased by optimizing k which is the ratio between the parts forming the geometry.

CST Microwave Studio is used to analyze the transmission and reflection response of the examined FSS and unit cell boundary conditions are used in order to getting periodicity for both directions. The simulation results of the transmission and reflection characteristics of the designed unit cell are demonstrated in Fig. 3. Obtained frequency response shows that the designed structure exhibits band-stop characteristics in dual frequencies. The first resonant frequency is at $f = 8.8$ GHz, and the second resonant frequency is at $f_2 = 11.2$ GHz. The lower and higher resonant frequencies can be controlled separately by adjusting the lengths and proportion of lengths.

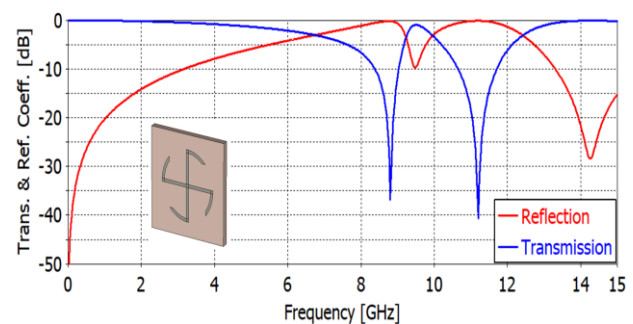


Figure 3. Transmittance and reflectance of the proposed unit cell illuminated by a normal incidence

Parametric analyzes are performed to investigate the effect of the angular variation of the metallic elements at the bottom surface on the transmission characteristic. Metallic elements at the bottom surface are positioned at 0 degree, 30 degrees and 40 degrees relative to metallic elements of the top surface. In Fig. 4, 0 degree, 30 degrees and 40 degrees configurations are demonstrated.

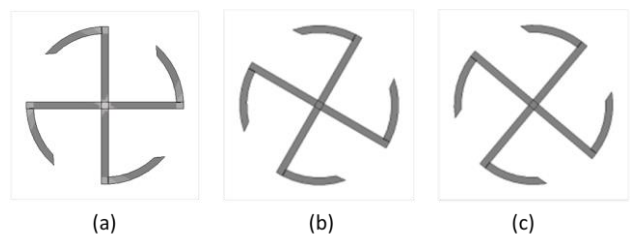


Figure 4. Angled configurations of the metallic elements at the bottom surface of the unit cells (a) 0 degree configuration, (b) 30 degrees configuration, (c) 40 degrees configuration

In Fig. 5, parametric study of simulation results of transmittance of the designed FSS at normal incidence for TE polarization is shown. It is clearly observed from the Fig. 5 that the angular configuration affects the operating frequencies and space between these frequencies.

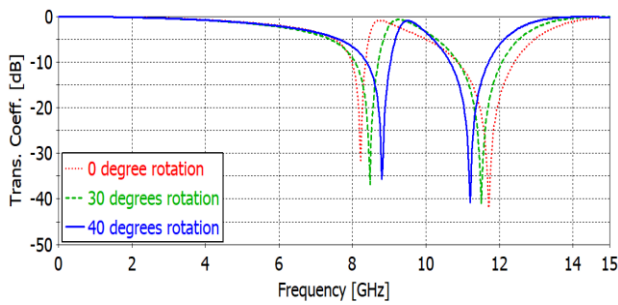


Figure 5. . Simulation results of angle-dependent transmittance of the 0 degree, 30 degrees and 40 degrees configurations of the metallic elements in the bottom surface

In Fig. 6. transmission response of the presented FSS for different incidence of EM wave. Angularly stable frequency characteristics is observed for up to $\theta=60$ degrees of several incident angle when $\phi=0$ for Transverse Electric (TE) polarization.

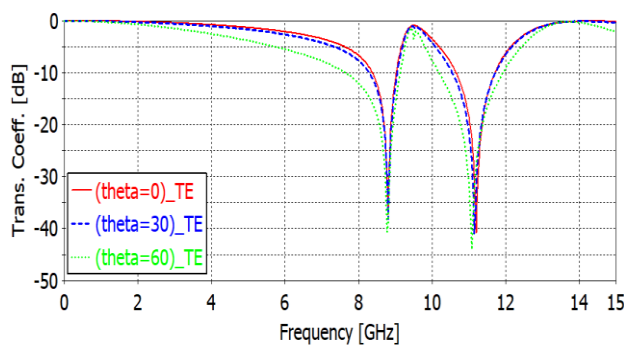


Figure 6. Simulation results of transmittance of the designed FSS illuminated by a different incidence for TE polarization

Metallic lines printed on front surface and back surface of the unit cell have four symmetrical arms. These symmetrical arms can diminish the sensitivity to the various incident angles.

The Designed structure maintains angular stability up to $\theta=60$ degrees of several incident angle when $\phi=0$ for Transverse Magnetic (TM) polarization. Stable frequency performance for TM polarization is observed from the Fig. 7.

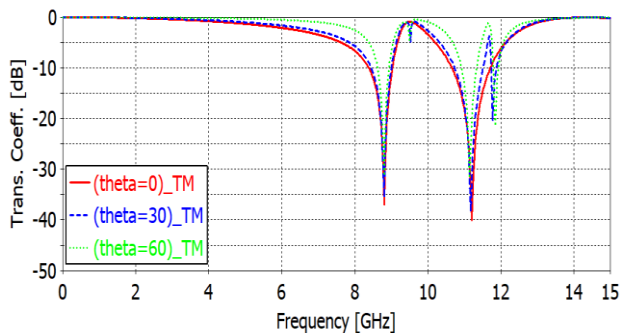


Figure 7. Simulation results of transmittance of the designed FSS illuminated by a different incidence for TM polarization

The surface current distributions of the presented unitcell at lower and higher operational frequencies for both TE and TM polarizations under normal incidence are demonstrated in Fig. 8 to express resonating parts.

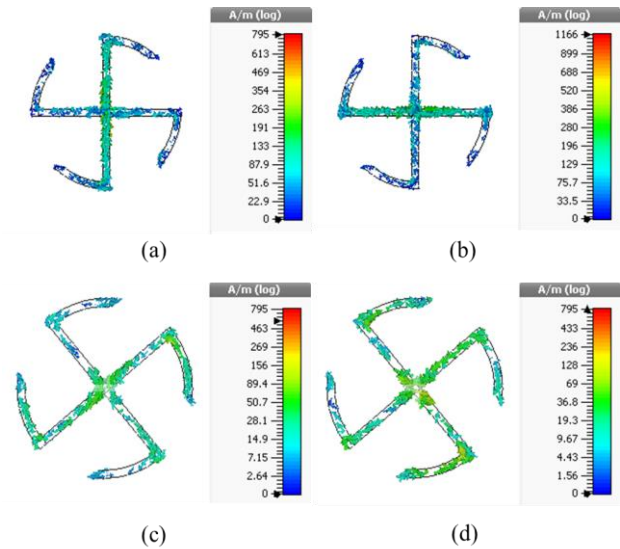


Figure 8. Surface current distributions of the designed FSS under normal incidence at (a) 8.8 GHz for TE polarization, (b) 8.8 GHz for TM polarization, (c) 11.2 GHz for TE polarization, (d) 11.2 GHz for TM polarization

III. CONCLUSIONS

A single layer closely located dual-band FSS is designed and simulated. Simulation results of the X band closely located FSS structure are obtained using CST Microwave Studio. The presented FSS is operating at 8.8 GHz and 11.2 GHz. The designed FSS with simple geometry have advantages such as low ratio of two resonating frequencies with 1.27s and ultra-thin design led to low mass. Also the presented design exhibits dual-band reject filter response. The simulation results show that the designed single layer structure provides a good frequency sensitivity up to 60 degrees of incident angles for TE polarization.

REFERENCES

- [1] S. Cimen, "A novel closely-spaced planar dual-band frequency selective surface," *Microw., Antennas Propag.*, vol. 7, no. 11, pp. 894–899, Aug. 2013.
- [2] B.A. Munk, "Frequency selective surfaces: theory and design" Wiley-Interscience, New York, NY, USA, 2000.
- [3] M. Yan, S. Q. J. Wang, M. Feng, W. Wang, C. Xu, Z. Li, L. Zheng, and H. Zhou, "A novel miniaturized dual-stop-band FSS for Wi-Fi application", in *2016 Progress In Electromagnetic Research Symposium (PIERS)*, Shanghai, China, 8–11 August, pp. 3447–3450.
- [4] S. N. Azemi, K. Ghorbani, and W. S. T. Rowe, "Angularly stable frequency selective surface with miniaturized unitcell," *IEEE Microw. Wireless Compon. Lett.*, vol. 25, no. 7, pp. 454–456, Jul. 2015.
- [5] G. H. Schennum, "Frequency-selective surfaces for multiple frequency antennas," *Microw. J.*, vol. 16, pp. 55–57, May 1973.

- [6] H. Fabian-Gongora, A.E. Martynyuk, J. Rodriguez-Cuevas and J.I. Martinez-Lopez "Closely spaced tri-band frequency selective surfaces based on split ring slots", *Electronics Letters*, vol. 52, no. 9, pp. 727–729, 28th April 2016.
- [7] R. Sivasamy and M. Kanagasabai, "A novel dual-band angular independent FSS with closely spaced frequency response", *IEEE Microw. Wireless Compon. Lett.*, vol. 25, no. 5, pp. 298–300, May 2015.
- [8] S. Unaldi, S. Cimen, G. Cakir, and U. E. Ayten, "A novel dual-band ultrathin FSS with closely settled frequency response," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 1381–1384, 2017.
- [9] S. Ghosh, and K. V. Srivastava, "An angularly stable dual-band FSS with closely spaced resonances using miniaturized unit cell", *IEEE Microw. Wireless Compon. Lett.*, vol. 27, no. 3, pp. 218–220, March 2017.
- [10] C.-N. Chiu, and W.-Y. Wang, "A Dual-Frequency Miniaturized-Element FSS With Closely Located Resonances", *IEEE Antennas and Wireless Propagation Letters*, vol. 12, pp. 163–165, February 2013.
- [11] T.K. Wu, "Four-band frequency selective surface with double square loop patch elements", *IEEE Trans. Antennas Propag.*, vol. 42, no. 12, pp. 1659-1663, 1994.
- [12] T.-K. Wu, S.-W. Lee, "Multiband frequency surface with multiring patch elements", *IEEE Trans. Antennas Propag.*, vol. 42, no. 11, pp. 1484-1490, 1994.

Automated Context and Text Analytics by Applying Cognitive Language Processing Tools

Ambruzs Csaba, Herczeg Dominik, Dobos Zoltán, Dr. Hajnal Éva

Óbuda University, Alba Regia Technical Faculty

ambruzs.95@gmail.com

herczdom96@gmail.com

hajnal.eva@amk.uni-obuda.hu

Abstract—IBM Watson [1] is a Cognitive system, what can analyze and interpret data - similarly to a human being - including unstructured text, images, audio and video, can learn and reason. This system has several capabilities (accessible through interfaces), which allow to execute context based natural language processing and interpretation. The purpose of the project is to provide support for human by significantly minimizing the effort, which is currently needed to analyze and understand large volume of unstructured audit text. IBM Watson gives an opportunity to identify IT risk factors, and compliance problems automatically, finding trends, and provide solution to improve problem detection while it helps to decrease the faults from manual human processing and improves efficiency.

In this article, the authors propose a method, which utilizes advanced natural language processing by using cognitive systems. Need to emphasize that the activity had to be carried out in an IT service management specific language area. The possible structures of dictionaries were investigated to adapt best the natural language processing capabilities and the required categorization, also the necessary pre-processing actions were reviewed. Within the dictionaries hierarchical mapping of the categorization levels (related to the IT risk and compliance area) is built up. Furthermore, the optimal combination related to the usage of nouns and verbs is determined to achieve higher hit ratio.

I. INTRODUCTION

Our future is the automated world[2], so we are trying to automate processes in many areas of the life. Automated machines, robots are used to reduce failure of the human working, and moreover to make the work easier and faster. There are two different approaches, the first one is the software agent and the second one is the hardware agent.

There are certain areas, which are highly automatable as they have many repetitive tasks, like IT services, which can be modelled into small executional parts. Similarly to other industries IT services or service management is also very cost sensitive, which is also a good driver for running automation

and serves like a good business need. Automation not only makes several repetitive tasks to be easier for us, but it means that employees can focus on tasks with more added values. This increases the success of a company.

The project what we are working on is a process automation. With the help of different text analytics tool, we have made a previously manual process faster, progressive, and more continuous. In this case our main goal is an automated text evaluation on a specific lingual area.

Solution selected for this task is a Watson based solution, what can analyze, interpret unstructured texts. These functions make it possible to apply for our task.

Similar solutions are being used in everyday life, but we are not thinking about how they work. The plainest examples: library and web searchers, these examples operating on the same principles, which is a keyword, apart from the algorithm, of course. Methods are searching for patterns, what comes from the searching criteria. This an efficient method while we are not looking for the context of the information along the interpretation

In this paper, the following section introduces data and the proposed method, Section 3 shows you how to create the model, Section 4 is about the results, Section 5 discusses and conclusions, Section 6 introduce the limitations of the solution, also introduce the planned future works, and finally section 7 say thank you for supporting our project.

II. DATA AND METHODS

A. Data to Analyze

The problems revealed during the audits, the related additional information, conditions and the descriptions of systems covers the data which are involved in the processing. This specific data is restricted to IT areas and within our project; it is limited to English text processing. The data are not structured, as they have free text nature. During processing what makes the challenge to be more complicated, that the concerns identified through the audit process are described in compound sentences or a single paragraph might be referring to several problem areas. Because of these, it is required to define a splitting method based on rules, and the analysis must run on these data for the efficient processing.

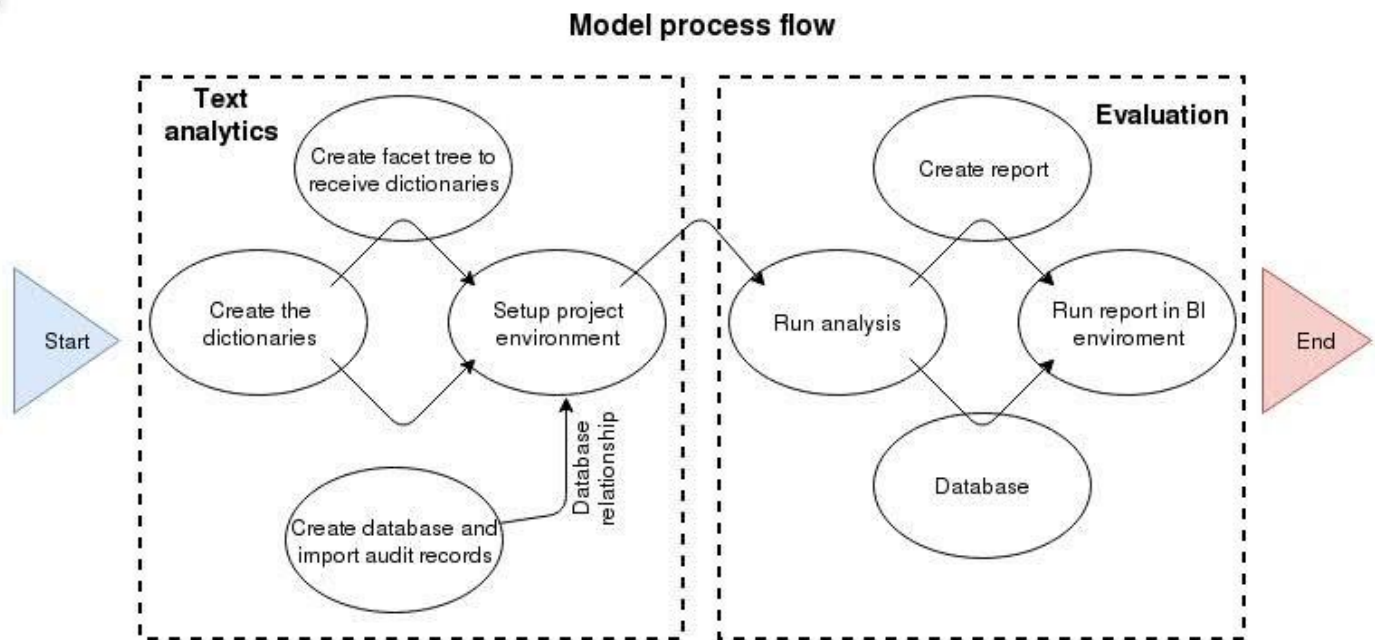


Figure 1. Overview. Shows the steps: Create Dictionaries, Export Pipeline, Run analysis, Run report based on our statistical model

The purpose of the analysis is to interpret these audit data, categorize texts and identify the risk issues in an automated way. As mentioned earlier this was done manually.

B. An overview of the model

The task is to create a process or model that can interpret texts in the described specific language area. The specialty given by two things, first text cannot be interpreted based on normal daily speech, second this IT service related language area is continuously improving and changing. We need to think of a solution that can support these specialties.

During the planning, we looked at some alternatives, and selected two of which we wanted to deal with:

- Traditional language analytics based solutions (key words, dictionary-based)
- Contextual based analytics (machine learning, context analysis)

It has been decided that both solutions will be analyzed and implemented, but as the first step the keyword driven analysis has been chosen to start with, considering that it will provide lots of experiences about the specific linguistic forms, which also required for the next step to apply machine learning and contextual analysis. The topic of the article is to introduce the approach applying traditional language analytics.

The figure 1 shows the overview of the process model: It is quite important to be familiar with this specific language area to be capable to prepare our processes for linguistic interpretation. The essence of the process is to identify important, meaningful words from the audit texts, which can be used for interpretation in this given linguistic area for

making decision. The purpose of the decision is to define what problematic area is described by the analyzed sentence in terms of IT risk and compliance. The decision is made on the bases of majority principle, which means to determine those keywords in the sentence, which are fitting the best for a known IT risk and compliance area.

The approach assumes that the sentence - describing a given problem - mainly contains specific terms for that area. There are exceptions like complex sentences, which are focusing on more than one problem or other sentences, which contains enumerations. Such sentences require data cleansing, data structure interpretation. The key in our procedure is to define this majority principle. Therefore, it is required to define first the environment, where these sentences, problem descriptions are analyzed, so this environment can be considered as a linguistic space for interpretation. This interpretation space can be mapped by a hierarchy, because the IT risk and compliance areas themselves are also mapped into specific categories, which can be divided further into subcategories along a hierarchy tree.

Our interpretation space contains 14 main categories and most of the categories contain 4 levels of subcategories. The success criterion of this process is to be able to map and associate the analyzed sentences into one specific element of the given category of this interpretation space. In the output, any sentence that corresponds to an element of the hierarchy is meaningful to us, in all other cases we are confronted a sentence what is not fitting in the given language area. Of course, in order to have a perfect decision, we need to cover this interpretation space perfectly, thus it is mandatory to

know the terms within this specific area. Since this is constantly evolving, the goodness of the model depends on how deep we can cover the space. Thus it can be stated that our model as more accurate and efficient as more text interpretations is executed.

Due to the fact that a particular word can fit into more than one specific point of the hierarchy, since the meaning of a given word is impacted by other words, thus we need to think of a solution that can handle not only a single word, but also word sequences or meaningful phrases. Regarding categorization, it can be observed that the main categories are mostly nouns, while subcategories are rather verbs. Illustrated on an example (Mgmt = Management)

“Approximately 39,667 privileged userids were not revalidated during 1Q 2017 continuous business need and privilege access revalidation activities.”

Main category:

“access” = noun -->Identify and Access Mgmt

Category 1:

“privileged userids” = noun --> Privileged ID Mgmt

Category 2:

“not revalidated” = verb -->No revalidation

The exact finding:

```

└ Identify and Access Mgmt
  └ Privileged ID Mgmt
    └ No revalidation
    
```

However, in the description of the procedure, stay on a frequency-based decision [3], which is performed as follows:

In the first step, we have to count the number of main categories for the sentence, and then summarize the same(1):

$$\begin{aligned} & \text{A} = \text{count}(\text{distinct}(\text{main}) \text{ for sentences}) \\ & \text{main} = \text{sum}(\text{A}) \text{ for main} \end{aligned} \quad (1)$$

Then calculate the total number of main categories in one sentence(2):

$$\begin{aligned} & \text{B} = \text{count}(\text{distinct}(\text{main}) \text{ for sentences}) \\ & \text{all main} = \text{sum}(\text{B}) \text{ for sentences} \end{aligned} \quad (2)$$

After these steps count a percentage of the main categories for sentences (3):

$$\text{main\%} = \text{main} / \text{all main} * 100 \quad (3)$$

With repeating the same steps doing this on category level 1-2-3, making sure that the parent changes each time because of the hierarchy of multiple levels of the tree (4):

$$\begin{aligned} & \text{A} = \text{count}(\text{distinct}(\text{cat1}) \text{ for main}) \\ & \text{cat1} = \text{sum}(\text{A}) \text{ for cat1} \\ & \text{B} = \text{count}(\text{distinct}(\text{cat1}) \text{ for main}) \\ & \text{all cat1} = \text{sum}(\text{B}) \text{ for main} \\ & \text{cat1\%} = \text{cat1} / \text{all cat1} * 100 \end{aligned}$$

$$\begin{aligned} & \text{A} = \text{count}(\text{distinct}(\text{cat2}) \text{ for cat1}) \\ & \text{cat2} = \text{sum}(\text{A}) \text{ for cat2} \\ & \text{B} = \text{count}(\text{distinct}(\text{cat2}) \text{ for cat1}) \\ & \text{all cat2} = \text{sum}(\text{B}) \text{ for cat1} \\ & \text{cat2\%} = \text{cat2} / \text{all cat2} * 100 \end{aligned}$$

$$\begin{aligned} & \text{A} = \text{count}(\text{distinct}(\text{cat3}) \text{ for cat2}) \\ & \text{cat3} = \text{sum}(\text{A}) \text{ for cat3} \\ & \text{B} = \text{count}(\text{distinct}(\text{cat3}) \text{ for cat2}) \\ & \text{all cat3} = \text{sum}(\text{B}) \text{ for cat2} \\ & \text{cat3\%} = \text{cat3} / \text{all cat3} * 100 \end{aligned} \quad (4)$$

If these results are available, then main%(own choice) is cut above a threshold to decrease the multitude of findings(5):

$$\text{Where}(\text{main\%} > 50) \quad (5)$$

Thus a list is created with a reduced probability, but in many case it contains more than one hit per sentence, because of the multi-level structure (several subcategories belong to a major category). To overcome this problem, weighted percentages are calculated for each level. Based on our tests, the higher levels reach a more accurate hit, so they get the highest weight, down the levels these weights decrease (6):

$$\begin{aligned} & \text{If}(\text{main\% is not null}) \text{ then}(\text{main\%} * 1000) \\ & \quad \text{else}(0) \\ & \text{If}(\text{cat1\% is not null}) \text{ then}(\text{cat1\%} * 100) \\ & \quad \text{else}(0) \\ & \text{If}(\text{cat2\% is not null}) \text{ then}(\text{cat2\%} * 10) \\ & \quad \text{else}(0) \\ & \text{If}(\text{cat3\% is not null}) \text{ then}(\text{cat3\%}) \\ & \quad \text{else}(0) \end{aligned} \quad (6)$$

Then sum these scores for each row (7):

$$\text{Sum} = \text{sum}(\text{main\%}, \text{cat1\%}, \text{cat2\%}, \text{cat3\%}) \quad (7)$$

In the last step the highest score is picked up for each sentence, so we reach our goal what is to identify only one finding for one sentence (8):

$$\begin{aligned} & \text{sort by} (\text{sentences}, \text{Sum}) \\ & \text{running-count} (\text{sum for sentences}) \\ & \text{count} < 2 \end{aligned} \quad (8)$$

III. CREATING THE MODEL

First step is to define data what we have to process. It is mentioned previously that data are audit records. Creating a database with specific parameters is needed for data storage. After that data have to be imported for processing from a centralized system.

IBM DB2 – database 2- is used to store data. After the installation we needed to create a non-default buffer pool[4] - cache - for reason, storing big volume of data. This solution is used to make our process faster.

In case of DB2, the objects of the relational database are organized into sets called schemas. A schema is a collection of named objects that provides a logical classification of objects in the database. So next step is creating a new schema based on the special buffer pool, to provide access to a bigger cache.

To store data, we had to create an intermediate table. Temporary data storage is not requirement in a complex database architecture in our case, so we created only one table for storage. Subsequently a star schema is used for making report, similar like this:

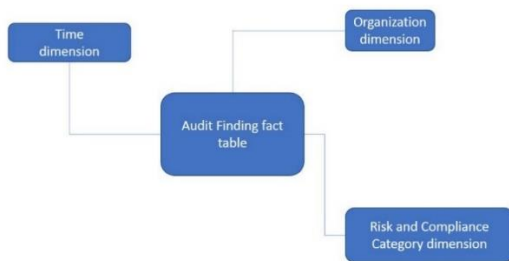


Figure 2. Star database schema

The table contains an ID with automate generate function yet, it is used for identifying and counting records. Table also contains additional information: Account name, Organization, Location, Date, Assessment, Tittle, Relative Size. These columns are stored in text format, and they contain our most information. Concern, Conditions are also stored in text format.

After creating the table, we have to upload it with data. It was solved with pull method, that imports data into our table from a centralized database. It makes capable to use this table to communicate with the natural language processing tool.

Next step is creating a dictionary, it is the most critical movement in our process. Analysis is based on keywords, so it needs to fill extensively. Watson Analytics Studio [1] capable to make unique dictionaries. To cover IT risk specific problem space, needed to design an efficient dictionary hierarchy. Our system needs to run along this IT risk hierarchy, to achieve fully qualified finding. While creating the dictionary we have

to pay attention, that the specified dictionary contains only category specified words, otherwise it would result fake findings. At first, by analyzing the existing audit records, we searched for words fitting into categories, and after reviewing those words, they were added into dictionaries. After loading, the measurement gives us a feedback about our work. By repeating these steps, we derived all of the dictionaries, what we can.

System allows us to create a “custom annotator”. But what does it mean? It means we could insert our IT risk dictionaries into an annotator - pipeline -, in addition to the basic English dictionary. The difference between the two kinds of dictionaries is that basic one aims to grammatically analyze a sentence and identify the parts of the text, while our dictionaries aim to identify the specific category. This feature allows us to categorize the texts. To continue the process, another tools (Watson Explorer Content Analytics, Cognos BI Report Studio) have to be involved into the solution work flow. This is needed because the Studio cannot give opportunity to export the results into relational database for finalizing the report.

Watson Explorer Content Analytics[1] - WEX - collects and analyzes structured and unstructured content in documents, email, databases, websites, and other enterprise repositories. By providing a platform for crawling and importing content, parsing and analyzing content, and creating a searchable index, Watson Explorer Content Analytics helps you perform text analytics across all data in your enterprise and makes that data available for analysis and search. Working with the two applications there is a way to connect them together. WEX enables to import our “custom annotator”, for use it for analyzing. In the WEX firstly we have to create an empty project. In the WEX similarly to Watson Analytics Studio we have to create a hierarchy to map the dictionaries, it is called “Facet tree”. It is important to create a good hierarchy, because the software can recognize the parent-children relations. It is important for us to “draw” the fully qualified branch, in the lower levels.

Last step is exporting the result into a relational database, what a report system, in this Cognos BI can use. As the results of the custom annotation gives us more than one category for a given sentence, we need to select the best candidates as finding. That we achieve by using the earlier mentioned statistics model - weighted percentage -.

IV. RESULTS

In this section, we are presenting the result achieved by our solution. Test cases were based on existing audit records, with human validation. Our database table is used to provide data for text analytics. The figure 5, 6 visualize our report quality, based on the previous points.

Before presenting the final results, we introduce how the system works, and the intermediate steps what format the final results. Let's see an example (Table I):

„Root cause analysis was not always performed upon a service level failure. Further, root cause analysis when performed did not always identify the actual trigger of a failure.”

Findings are the following words: identify, identify, was not always performed, root cause, root cause analysis, Root cause analysis

TABLE I

RESULT OF AN EXAMPLE. FIRST COLUMN IS THE FOUND KEYWORD, SECOND IS THE TARGETED MAIN CATEGORY, AND THIRD IS THE FOUND CATEGORY. THE FALSE TRUE ANSWER IS BOLD>.

Keyword	CategoryMain	Founded category
identify	IT_Risk_Management_Services	Exception_not_identified
identify	IT_Risk_Management_Services	RCA_inaccurate_or_incomplete
was not always performed	IT_Risk_Management_Services	RCA
root cause	IT_Risk_Management_Services	RCA
root cause analysis	IT_Risk_Management_Services	RCA
Root cause analysis	IT_Risk_Management_Services	RCA
Overall	6	1:5(false:true)

Figure 3 visualizing the Table I.

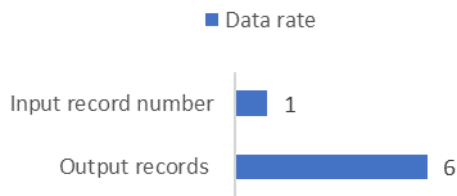
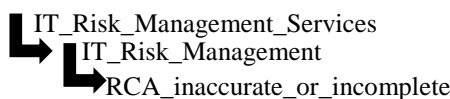


Figure 3. Findings per record

Word repetition is discoverable, because a word can belong to more than one category. To solve this problem, we need to apply a report system as we mentioned. Let see the “identify” word. This belong to two categories, however in the first case it reached only 16%, in the second case 84%. This difference is due to the other words which strengthens this branch for example: “root cause analysis”, so the final vote is based on the 2nd “identify”:



This solution ensures that finally only one result is displayed on the output. The test running on 97 records, then

with summarizing the data we get the following result (Figure 4).

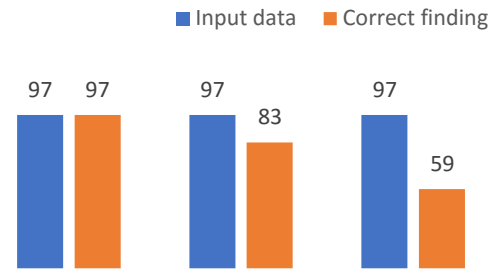


Figure 4. Correct Conversation Rate

In a percentage format, it looks like this (Figure 5), the correct finding percentage is significantly decreasing with the category level:

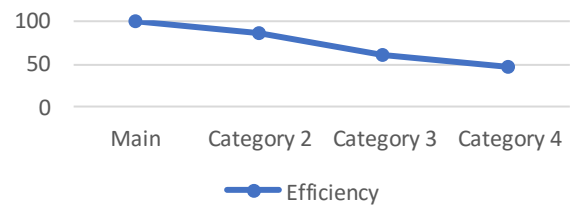


Figure 5. Percent hit rate

Summing up the results with hundreds of audit records, we get the Table II.

The upper blue part of the table shows a good ratio while the lower parts are getting weaker - the weakness can be attributed to the non-sufficient number of audit records what causes that the dictionary not filled up satisfactorily - but overall, despite the weakness of the lower part of the table, the average accuracy of 83.3 was achieved.

TABLE II : SUMMARIZED RESULT FOR 574 RECORD

Account	Accuracy	Manual	WEX
Account_1	100,0%	15	15
Account_2	96,2%	26	25
Account_3	94,7%	38	36
Account_4	94,4%	18	17
Account_5	93,8%	16	15
Account_6	92,7%	55	51
Account_7	91,7%	12	11
Account_8	90,9%	11	10
Account_9	90,5%	42	38
Account_10	89,3%	28	25
Account_11	87,5%	16	14
Account_12	87,5%	24	21
Account_13	87,5%	8	7
Account_14	85,7%	7	6
Account_15	85,0%	20	17
Account_16	83,9%	31	26
Account_17	81,5%	27	22
Account_18	77,8%	18	14
Account_19	76,0%	25	19
Account_20	68,2%	66	45
Account_21	68,0%	25	17
Account_22	66,7%	18	12
Account_23	66,7%	9	6
Account_24	47,4%	19	9
Grand Total	83,3%	574	478

V. CONCLUSION AND DISCUSSION

In this paper a complex text analyses was presented with its result. In summary the results are satisfactory.

Based on our model and approach, we have reached a relatively high hit ratio and accuracy by applying this language analytics model. Accuracy means in our case that this solution is able to identify the correct main category above 90% and in the second category the accuracy varies between 65-85%. This level of accuracy is good base for automation and decrease the level of manual effort. We still can identify space for improvements, which we can approach from two aspects. First aspect is the goodness of the dictionary itself, here we need external support having good level of knowledge about this specific area.

However it can be seen that at higher hierarchy levels - the top levels- we can achieve a flawless or almost completely flawed hit rate (fig 5). Towards the lower levels, this ratio was decreasing, which can be attributed to the degradation of processes in many branches, and the diversity of the content of the verb report. The weakness of the lower levels can be attributed also to the dictionary's weaknesses, which means that in those dictionaries there are not enough supporting words.

All test results were discussed concerning the success of the analysis, and it was concluded that in the further text analysis the improvement of the hit rate can be approached from two sides. First, let's take the easiest one to refine the dictionary's content. In this case, our knowledge about this specific area is not appropriate, so it is required to involve an external help - a person who had made manual analysis earlier - to refine the dictionary. We have also tried to consider the structure of the hierarchy when uploading the dictionary. The upper two levels of the hierarchy point to the definition of the subject, so in this case the given levels are filled with nouns defining the fundamental problem. The lower levels under each major categories refers to some parts of the processes, so it is better to moderate the number of adjectives, nouns in the dictionaries at these levels, but use rather verbs to create accurate results. This plays an important role because, if the dictionary is populated by this way, the approach will find more meaningful words for the higher level and for the lower levels, so the statistical model will give higher weight than any possible false results. Second, clearing the text can move to the desired direction. Input data contains a lot of specific characters - "; ' etc." that are removed to make the device work optimally.

In addition, the text contains many non-noticeable control characters - line breaks, etc. - which also have an affecting feature for the analysis; these are also removed from the text in some cases. As the last step of the content purification, it is necessary to mention that an input data consists of several sentences in most cases. These sentences are usually not describing only one fault, so it is necessary to analyze these sentences separately and to display them as separate texts between the output data, keeping in mind, of course, the related original text.

VI. LIMITATIONS AND FUTURE WORKS

Knowing that the tool is key word-based, so it looks for matching pattern in the text, it propounds various problems, the most significant of these being the spelling error. In the knowledge of the problem, we can put up a similar problem - a foreign word embedded in the word connection -, which means that we will not be able to recognize the specimen in this case, so after detecting these problems, we started looking for another similar text analytical method, and this is Watson's Cognitive Tool what is able to demonstrate the ability to use machine learning by training through number of samples.

Keeping in mind the success of our project in the following, we try to recognize this cognitive solution and increase the categorization goodness in this special language area. The cognitive path gives two paths, first one is also a keyword-based system that, with the help of the applied dictionaries, is able to further refine system by machine learning. Another one is a new one, which does not require any previously created dictionaries, it is based on machine learning only, of course, a sufficient number of samples is required, and then the approach will give us automatically the highest probability data.

VII. ACKNOWLEDGMENT



SUPPORTED BY THE ÚNKP-16-1 NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

We acknowledge the financial support of this work by the Ministry of Human Capacities of Hungary

REFERENCES

- [1] W.-D. J. Zhu and International Business Machines Corporation. International Technical Support Organization., "IBM Watson Content Analytics discovering actionable insight from your content." 2014.
- [2] R. Berger, "Of robots and men - in logistics." 2016.
- [3] Dudás László, "Alkalmazott Mesterséges Intelligencia," 2011.
- [4] J. Z. Teng and R. A. Gumaer, "Managing IBM database 2 buffers to maximize performance," *IBM Syst. J.*, 1984.

NB-IoT Technology and its Application

Szabó Dávid^{*, **}, Mákl Roland^{*, **}, Dr. Nagyné Dr. Hajnal Éva^{*}

^{*} Óbuda University, Alba Regia Technical Faculty

^{**} Albacomp RI Ltd.

szdavid.opx@gmail.com

makl.roland@live.com

hajnal.eva@amk.uni-obuda.hu

Abstract— In June of 2016 the 3rd Generation Partnership Project (3GPP) completed the standardization of Narrow Band IoT (NB-IoT). NB-IoT is a Low Power Wide Area Network (LPWAN) radio technology for the Internet of Things using conventional cellular network. It focuses on increased coverage, lower costs, extended battery life and enabling a vast number of connected devices. Using unexploited LTE resources, narrow band technology does not require new infrastructure to be built. In this article, a universal NB-IoT device and software framework was designed. It can be equipped with several types of sensors – along with its platform. The final tests of the device can be performed according to the deployment of base stations.

I. INTRODUCTION

The fourth industrial revolution (Fig. 1) not only brings digitalization, robotization and automation, but establishes a different business paradigm which, with all its innovations raises up a new concept, called Industry 4.0 for short.

This revolution is still ongoing, more and more cyber-physical system appear that means informatics, mechanics and software are tightly coupled. Cyber-physical systems are realized by interconnections of embedded devices by means of wired, or more preferably by wireless technology. Problem arises when devices had to be connected to a network that is hardly accessible, or there is no way to use conventional methods to do so, or financially not viable.

These problems were addressed by 3rd Generation Partnership Project (3GPP) in June of 2016. In its Release 13 a new technology, called Narrow Band IoT, was introduced.

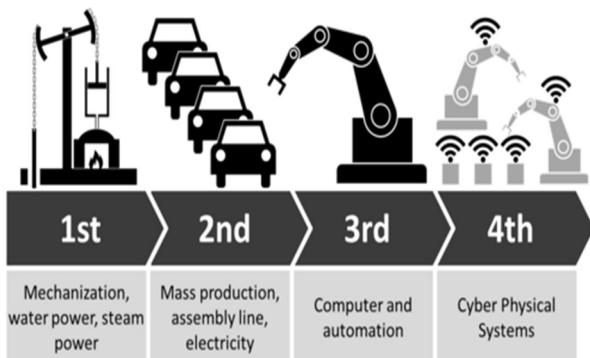


Figure 1. Industrial revolutions and future view

Narrow Band is a Low Power Wide Area (LPWA) standard radio technology, specifically designed for the Internet of Things. It focuses on improved coverage, reduced costs and prolonged lifetime (Fig. 2).

According to Telekom's analysis, approximately 3 billion LPWA devices are to be connected around the globe by 2023 [1].

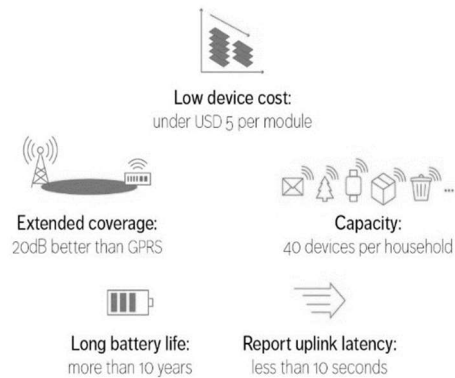


Figure 2. Benefits of NB-IoT [2]

That is why we are motivated to design a full NB-IoT solution, because currently there is no general solution yet.

II. TECHNICAL OVERVIEW

A. Low Power Wide Area

LPWA networks give reasonable range with minimal power consumption. According to Machina's research this will be the fastest growing IoT technology [3].

These networks have two main aspects:

1. Low Power

Devices designed for LPWA networks are optimized in hardware and software as well, thus able to function for many years powered from a single battery. Power consumption is so low, it is comparable to the self-discharge of the battery, making it unnecessary to replace the power source ever. Therefore, these devices can be placed in hard-to-reach environments.

2. Wide Area

Technological advances make it possible to have increased transmission power and receiver sensitivity, obstacles and interference are easier to cope with.

Conventional networks, like Bluetooth, Wi-Fi or ZigBee, along with classical cellular technologies, are not efficient enough, requires considerable amount of investment and power consumption is a concern as well.

On the other hand, LPWA devices are relatively cheap, uses existing infrastructure and require minimal, or no maintenance.

B. Narrow Band IoT

Table 1. Technical parameters

Frequency band	NB-IoT (LTE) FDD bands 1, 2, 3, 5, 8, 11, 12, 13, 17, 18, 19, 20, 25, 26, 28, 66, 70
Mode	Half-duplex FDD type B
MIMO	not supported
Bandwidth	180 kHz
Multiple Access	Downlink: OFDMA Uplink: SC-FDMA
Modulation	Downlink: QPSK Uplink: Single Tone: $\pi/4$ -QPSK, $\pi/2$ -BPSK Multi Tone: QPSK
Coverage	164 dB (+20 dB GPRS)
Data rate	~25kbps downlink and ~64 kbps uplink
Propagation	<10 seconds
Power savings	eDRX, Power Saving Mode

1. Channel scanning and connection establishment

The key feature of an NB-IoT device is the ability to adapt to changes in the environment. This property is defined in Physical Random Access Channel (PRACH) requirement which ensures sufficient link quality to cells [4].

According to 3GPP Release 13, a device, during its power-up, is scanning for available channels and tries to connect to them using one of the three signals [5]:

1. Narrowband Cell Reference Signal (NRS)
2. Narrowband Primary Synchronization Signal (NPSS)
3. Narrowband Secondary Synchronization Signal (NSSS)

NRS signal is used by the User Equipment (UE) to determine the performance of the downlink and is present in every downlink subframe.

NPSS and NSSS estimates (Fig. 3) time and frequency properties with a primary signal in every 5th subframe and a secondary signal in every 9th subframe of frames with even numbers.

With these properties at hand, the device is ready to receive Narrowband Physical Broadcast Channel (NPBCH) signal, in which, it acquires Master Information Block (MIB-NB). The MIB-NB informs the UE, in what modes the cell is operating [6]:

- a) Stand-alone (reused GSM frequency band)
- b) In-band (inside the LTE spectrum)
- c) Guard-band (next to an LTE Physical Resource Block - PRB)

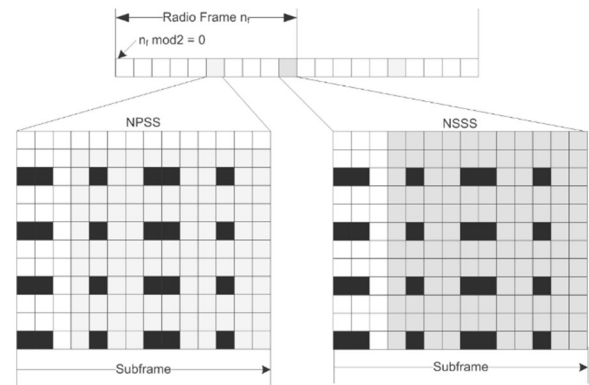


Figure 3. Primary and secondary synchronization signals [7]

Every NB-IoT device should support all modes and configure itself based on the information provided.

Narrow Band technology does not require as much resource as standard LTE or GSM, as datarate is lower, handover is not implemented and there is no support for Multiple-Input and Multiple-Output (MIMO) [8].

2. Energy Saving Methods

But the lifetime of an NB-IoT device is still heavily dependent on the power source it uses, therefore it is critical to optimize power consumption. Optimizations can be done at hardware level, but a thoroughly designed software is the essence.

3GPP Release 12 defines the Power Saving Mode (PSM), which enables a device to sleep indefinitely. In this state, the UE is unreachable but can be woken up by its internal components, or when the Tracking Area Update (TAU) times out (Fig. 4).

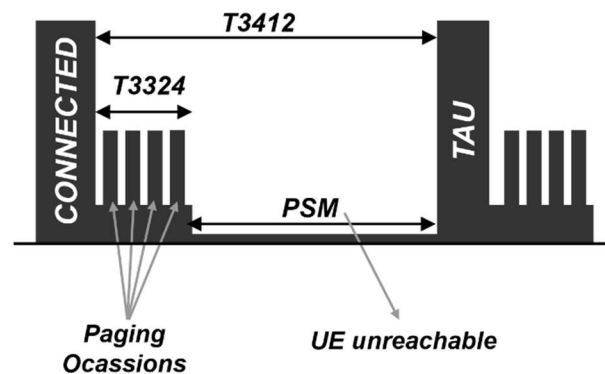


Figure 4. Release 12 Power Saving Mode [9]

3GPP Release 13 defines another power saving technique, the Extended Discontinuous Reception (eDRX), in which the UE can sleep for longer time before it checks back into the network (Fig. 5-6). The UE tells the network how many units of time it would like to sleep and during that time, traffic is queued.

eDRX is useful for applications when frequent downlink is expected.

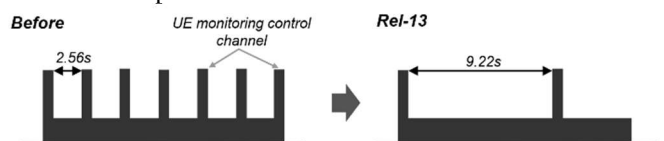


Figure 5. Rel-13 eDRX connected [9]

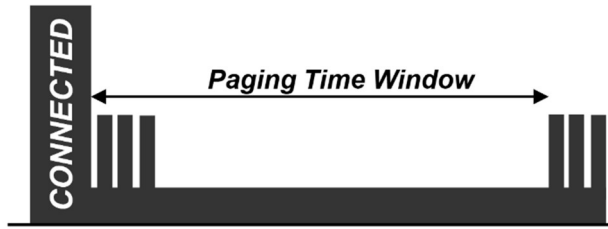


Figure 6. Rel-13 eDRX idle [9]

III. DESIGN

A. Specification and objectives

During design phase, the main motivation was to develop an NB-IoT device, which lets future adaptation and takes the integration realized easier.

To achieve this goal, we designed a hardware which not only complies with the IoT norms (minimal power consumption and size), but it can be used in many applications by providing sufficient interface for external components, like sensors and control elements.

B. Hardware

The hardware components were chosen to meet the requirements of technological trends to the greatest extent possible. Key considerations are functionality, power consumption and size. The device consists of the following main elements (Fig. 7):

1. NB-IoT enabled module
2. microcontroller
3. power source
4. peripherals and extensions

Nowadays Quectel and U-Blox has working solution for NB-IoT. Both modules are based on Huawei's chipset, but only the Quectel BC95 is available in our region [10].

Because the BC95 module has no dedicated user-application processor, we had to use an individual one in order to control the module, sensors and manage resources. We selected a modern, yet ultralow consumption type one from the range of STMicroelectronics ARM Cortex-M architecture processors. The STM32L041G4 was chosen due to its processing power, low-power modes, peripherals and package [11].

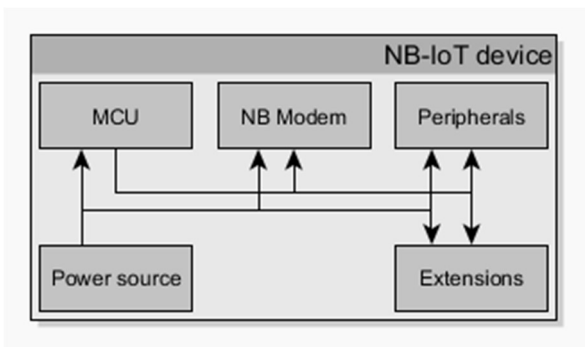


Figure 7. Architecture of the device

The built-in Real Time Clock (RTC) module can keep track of time while the whole device is in sleep mode.

As a matter of constant concern for our world, IT security, encrypted data can be easily generated by its AES-128 crypto-module.

Power is provided by a standard Li-Ion battery, whose capacity will be chosen based on preliminary power consumption and required lifetime.

Regarding the status of the microcontroller and the modem, two major and several minor states were distinguished in the calculations:

1. Active state
 - a. data acquisition, measurements
 - b. processing
 - c. transmission

2. Inactive state

In the active state, the microcontroller gathers data from the environment by reading sensors, process the acquired data and transmits them through the NB-IoT module. After completion, the whole device enters the inactive state.

While the device is inactive, power consumption is at minimum, microcontroller is stopped and the NB-IoT module is in PSM or eDRX. If any event happens, the device enters into the active state again and performs actions.

Time spent in the states and the actions performed are periodic and almost equal in power consumption, therefore lifetime can be calculated based on the individual power consumption of different processes. Calculations are based on electrical charge as a resource:

$$Q = \int I(t)dt = I_{avg} * \Delta t \quad (1)$$

Consumption of different states and actions can be summed up to get the amount of charge used in a period of time. Then, the sum of these charges and the time of the period gives the average consumption.

$$Q_i = I_i * t_i \quad (2)$$

$$Q = \sum_{i=0}^n Q_i = \sum_{i=0}^n I_i * t_i \quad (3)$$

$$I_{avg} = \frac{Q}{\Delta t} \quad (4)$$

Battery capacity can be calculated from the average current extended to the required lifetime and estimated efficiency:

$$Q_{bat} = I_{avg} * t_{lifetime} * \frac{1}{\eta} \quad (5)$$

For example, if a device has an average power consumption of 100mA for 10 seconds, then 10uA for the rest of the day, battery capacity for a 10 years lifetime with 75% efficiency would become:

$$Q_0 = 100mA * 10s = 1As \quad (6)$$

$$Q_1 = 10uA * 86390s = 0.8639As \quad (7)$$

$$I_{avg} = \frac{Q_0 + Q_1}{1 \text{ day}} = \frac{1As + 0.8639As}{86390s} = 21.58uA \quad (8)$$

$$Q_{bat} = 21.58uA * 10 \text{ years} * 75\% = 2520mAh \quad (9)$$

, which is not too large considering 10 years of operation. Of course, most of the applications require more active time than 10 seconds a day, but it can be admitted, it is possible to operate a device for many years from a single battery cell.

C. Platform

To make a robust solution and to facilitate future development, we decided to create a complex system, which we call platform, to give an abstraction to the hardware, software and web service as a whole. With this abstraction, new applications can be implemented without spending time on development.

Aside from minimizing the size of the design, we added extension connectors to support connection for common serial interface components, most commonly using I2C, SPI or UART. We dedicated these peripheral signals as well as some general I/O lines for control purposes.

For simplifying development of the device, a concept of a complete firmware library was created, in which all the functionality would be implemented and hidden from the programmer in the form of an API. This library would act like an operating system and would guarantee minimal power levels and robustness by design. The API would reduce application logic to a few function calls. Furthermore, a graphical programming language could be developed for this purpose. The library could handle Firmware Over-The-Air (FOTA) and implements failsafe mechanisms.

Data are transmitted to the web in the popular JSON format, which enables flexible usage, different types of data can be enclosed in a single message and almost every web service could handle JSON format.

Transmitted messages always contain unique identifiers, battery voltage level, network statistics and auxiliary flags, so a fleet management solution can be easily built to constantly monitor deployed devices, giving notifications on different events.

IV. CONCLUSION

NB-IoT has a great potential for the world today. By using the LTE network, it is cheap to implement by operators and can emerge rapidly. It provides good coverage and with proper design, can last for 10 years or for more.

With our design, many applications can be implemented by attaching a little extension board to the device and thanks to the platform, it requires minimal configuration to operate.

According to provisions, vast number of NB-IoT devices will be connected to the internet, which means, Big Data solutions or Artificial Intelligence should be implemented later to handle and process huge amount of data.

Some of the applications are focusing on smart devices used in cities and facilities:

A. Smart City

There is no better use of NB-IoT technology, where long-life and maintenance-free solutions are most critical. Smart city is an idea to manage urban assets based on IoT solutions.

B. Smart lighting

A city with many street lamps consumes a considerable amount of power. By replacing them to more efficient LED lights and remotely controlling them individually, lamps could be lit where it is needed, thus minimizing utility costs.

C. Smart parking

It is always a challenge to find parking place in a crowded parking lot. It takes time, fuel and is stressful. If every parking place had a sensor to indicate vacancy and provides connection to the internet with NB-IoT, vehicle owners could get notified about available spots nearby, or they could search them manually on their phone.

D. Smart Bin

Cities have their schedules for waste collection. This means excessive waste is not removed before the next schedule or there is no need to remove at all. Therefore, managing waste is not as efficient as it could be. Using NB-IoT, waste collection could be scheduled on demand, saving time and fuel.

The device and its platform significantly accelerates the development of further solutions. Now in Hungary, and surrounding countries had no available NB-IoT network, we could not manage the final tests of our design yet.

Based on provisions and preliminary calculations, carefully designed NB-IoT devices testify properties stated in 3GPP standards. Narrow Band IoT devices can be run maintenance-free throughout their lifetime due to their low power consumption.

ACKNOWLEDGMENT



SUPPORTED BY THE ÚNKP-16-1 NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

We acknowledge the financial support of this work by the Ministry of Human Capacities of Hungary

REFERENCES

- [1] Telekom, "Narrow-Band IoT: A network designed for the 'simple things' in life." [Online]. Available: <https://www.telekom.com/en/company/details/narrow-band-iot--a-network-designed-for-the--simple-things--in-life-363362>.
- [2] S. Landström, J. Bergström, E. Westerberg, and D.

- Hammarwall, “NB-IOT: A sustainable technology for connecting billions of devices,” *Ericsson Rev. (English Ed.*, vol. 93, no. 2, pp. 8–16, 2016.
- [3] GSMA, “GSMA Mobil IoT Initiatives | Low Power Wide Area Technology.” [Online]. Available: <https://www.gsma.com/iot/mobile-iot-initiative/>.
- [4] A. D. Zayas and P. Merino, “The 3GPP NB-IoT system architecture for the Internet of Things,” *2017 IEEE Int. Conf. Commun. Work. ICC Work. 2017*, pp. 277–282, 2017.
- [5] Y. P. E. Wang *et al.*, “A Primer on 3GPP Narrowband Internet of Things,” *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 117–123, 2017.
- [6] N. Mangalvedhe, R. Ratasuk, and A. Ghosh, “NB-IoT deployment study for low power wide area cellular IoT,” *IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, no. 1, 2016.
- [7] J. Schlien and D. Raddino, “Narrowband Internet of Things Whitepaper,” p. 42, 2016.
- [8] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, “NB-IoT system for M2M communication,” *2016 IEEE Wirel. Commun. Netw. Conf. Work. WCNCW 2016*, no. Wd5g, pp. 428–432, 2016.
- [9] M. Blanco and P. Oloriz, “A cellular technology connecting the Internet Of Things Agenda Why NB-IoT Technical Fundamentals Test Challenges Summary.”
- [10] Quectel, “Quectel BC95 NB-IoT Specification,” pp. 6–7.
- [11] ST, “STM32L041x4 STM32L041x6 Access line ultra-low-power 32-bit MCU ARM®-based Cortex®-M0+, up to 32KB Flash, 8KB SRAM, 1KB EEPROM, ADC, AES,” no. December, 2016.

Matplotlib in the geospatial analyzes

G. Nagy*

*Óbuda University, Alba Regia Technical Faculty, Institute of Geoinformatics
 nagy.gabor@amk.uni-obuda.hu

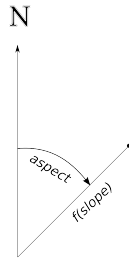


Figure 1. The polar coordinate system of the diagram.

Abstract—The diagrams are useful tools to illustrate the result of an analysis. In many cases the classical tools (the diagram creator modules in the spreadsheet softwares) can not make the imagined diagrams. For example diagram which was suggested in my previous article to illustrate the distribution of slope and aspect of an area. The Matplotlib is an Open-source plotting library, which can make varies diagrams with varied settings. This tool can be used well in spatial analyzes. Data processing and charting can be done in an Python program.

Index Terms—Matplotlib, Python, GIS

I. INTRODUCTION

The figures are very important parts of scientific articles and education materials. A lot of application can create various diagrams. Probably, the most famous and easy choice is the chart creator function of the spreadsheet softwares, but these tools provides only the basic chart types, and the customization is also limited.

This paper present an open source tool for this task with practical examples in the geospatial analyzes.

II. A VISUALIZATION TASK

The distribution of slope and aspect can be shown an polar diagram, where the angle is the aspect and the distance depends from the slope (in the following examples the distance is the square root of the slope). (See in Figure 1.) The distribution of the different areas are shown an scatter chart in this polar coordinate system. ([1].)

The distribution of the points in these scatter diagrams represents the distribution of the slope and aspect in an area. The [1] uses a program, which creates SVG files from the datasets.

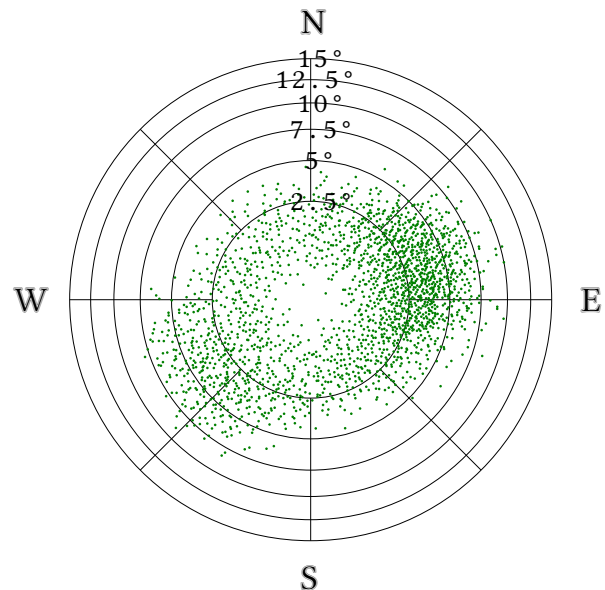


Figure 2. The scatter diagram created by an SVG based program.

III. THE MATPLOTLIB

The Python language [2] is very popular in the scientific projects [3], [4], many Python tools aid research works. The Matplotlib [5] is an open source Python module for creating diagrams.

The diagrams presented in the Section II can be created by the Matplotlib tools easier than the SVG-based solution. (See in Figure 3. and Figure 4.)

The Matplotlib based program is more clearly, because it need not create the XML text description of a lot of required parts of the SVG files.

IV. SURFACE VISUALIZATION IN MATPLOTLIB

Matplotlib can show surfaces by different methods. The surface data is stored in a 2 dimensional NumPy [6] array. This array follows the Matlab convention: the X coordinate depends from the column number and the Y coordinate depends from the row number. The Gdal [7] module can read raster data from varied file formats to a NumPy array.

The direction of the rows are reverse in the Gdal read arrays than the upper convention. The rows of the array can be reversed by the `flipud` function of the NumPy.

A. Colored maps

The Matplotlib can create colored map representation of the surfaces. The simplest way for this is

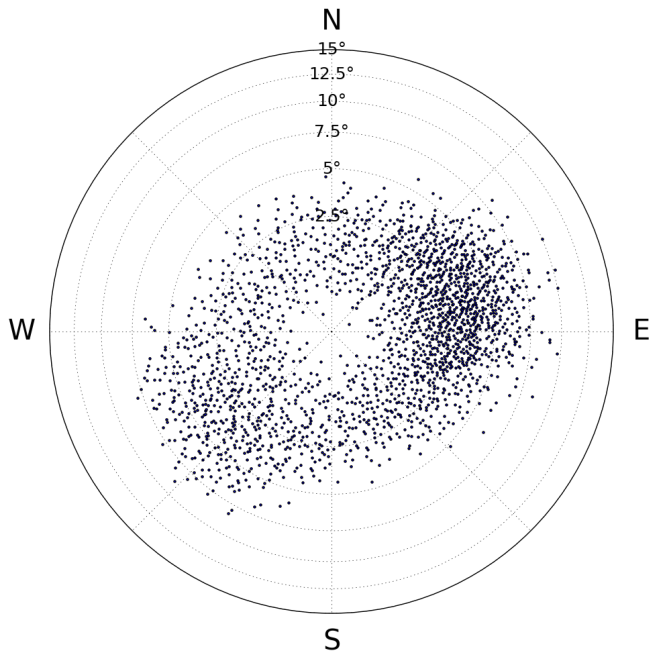


Figure 3. The scatter diagram created by an Matplotlib based program.

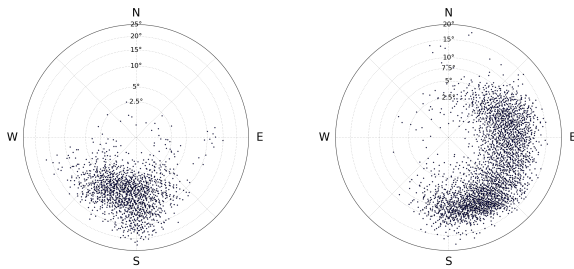


Figure 4. Other scatter diagrams created by the Matplotlib based program.

the `matplotlib.pyplot.pcolor()` function. (The `matplotlib.pyplot` usually is used by the `import matplotlib.pyplot as plt` command in the Python programs, and in this case this function can be used by the `plt.pcolor()` form. In the following, I use this form in the sample codes.) The `plt.pcolor(z)` creates a colored map from the `z` two-dimensional NumPy array by the default colormap, where `z` is a two-dimensional array with the elevations.

Other colormap can be used by an optional argument of the `plt.pcolor()` function. For example: `plt.pcolor(z, cmap=plt.cm.gist_earth)`. (See in the Figure 6. and the Figure 7.)

The Matplotlib can create shaded surface map. (See in the Figure 8. and the Figure 9.) An `LightSource` object configures the azimuth and the altitude of the light source of the shading.

B. Contour maps

The `plt.contour()` function creates contour map from the surface data. For example: `plt.contour(z)`. (See in the Figure 10.)

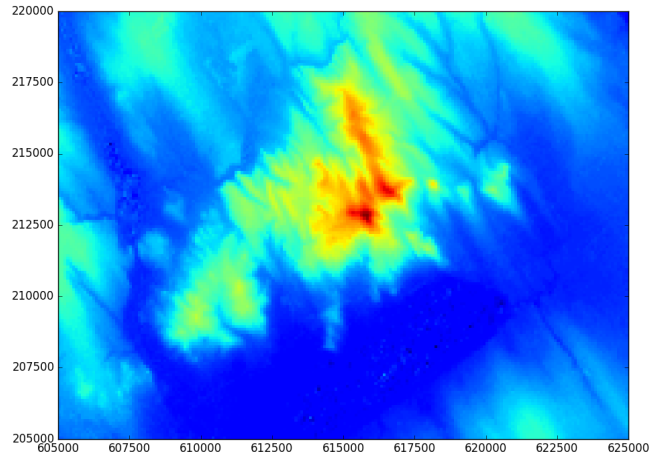


Figure 5. The surface of Velence Hills with the standard colormap of the Matplotlib

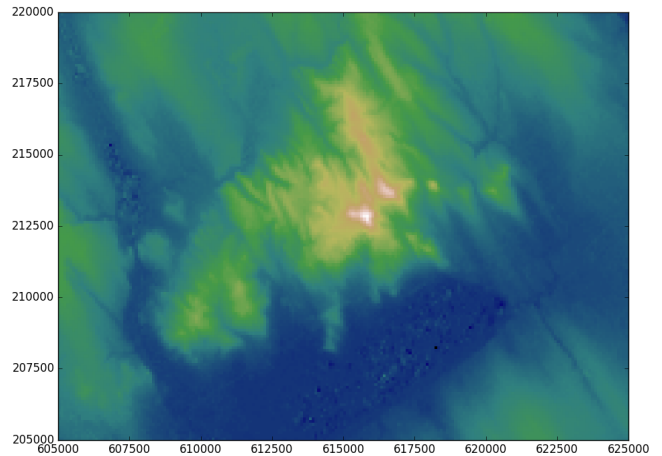


Figure 6. The surface of Velence Hills with the `gist_earth` colormap of the Matplotlib

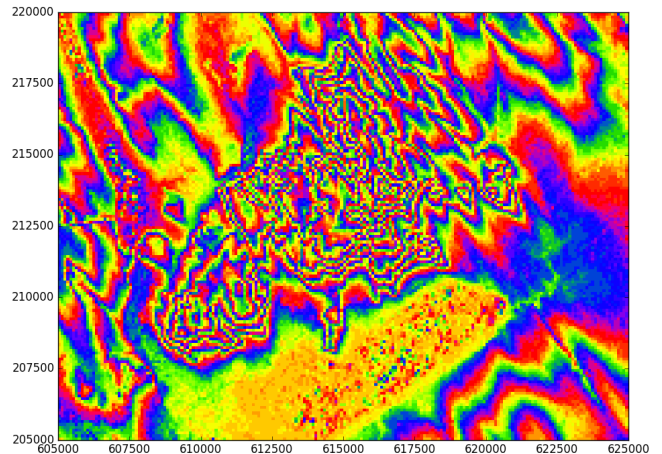


Figure 7. The surface of Velence Hills with the `prism` colormap of the Matplotlib

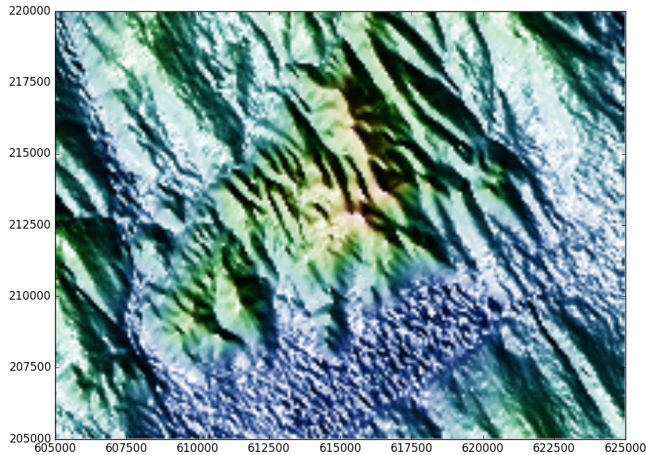


Figure 8. The surface of Velence Hills with shaded map by Matplotlib

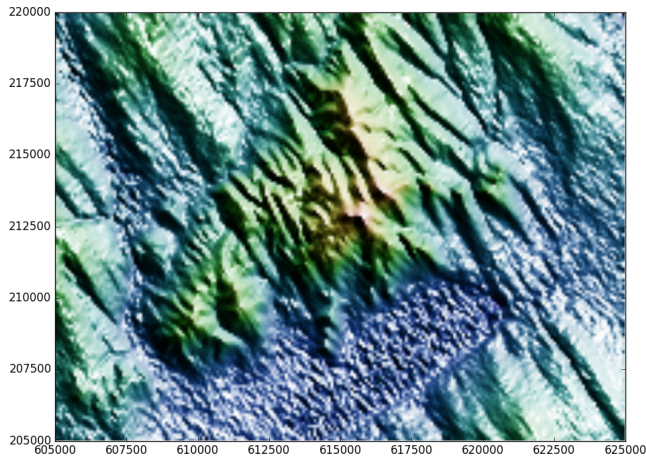


Figure 9. Shaded map with opposite light direction than the Figure 8.

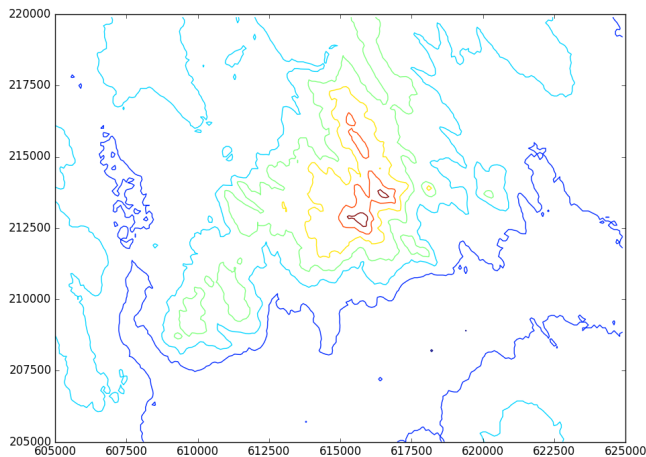


Figure 10. The surface of Velence Hills with contour lines

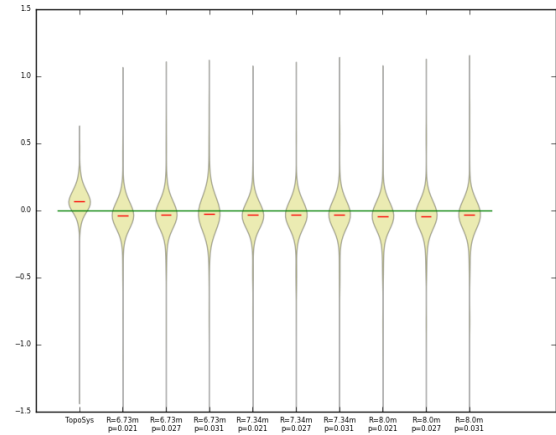


Figure 11. Violin plots of the distribution of LiDAR-based surfaces compared to geodetic survey.

The Matplotlib contour charts are not the typical contour maps. The program uses different colors for the different elevations. (The colormap can be changed.)

V. OTHER MATPLOTLIB APPLICATIONS

Matplotlib provides a lot of chart types. Any of them may be useful in a research.

For example LiDAR-based surfaces are comparable with geodetic surveys, the difference between the elevations from the surfaces and the geodetic survey can be calculated each point of the survey. The distribution of these values characterizes the method which created the surface from the LiDAR point cloud.

The violin plots [8] are good tools for representing distributions. The Matplotlib can create violin plot from a dataset.

A Matplotlib plot can contains more violins from different datasets. For example the first element of the Figure 11. and Figure 12. are the distribution of the errors (differences between the surface and the geodetic survey) in a surface created by an commercial LiDAR processing software (TopoSys), and the other violins shown same results by a new suggested LiDAR processing methods with different parameters (Denoted R and q).

Matplotlib based Python programs can create animated violin plots, one of the parameters can be changed in the time. This possibility are very useful, because an animation can show a lot of combinations of the R and q values.

VI. CONCLUSION

Matplotlib is an useful tool for creating scientific and educational charts. The diagrams can be created by Python programs. These programs also can be calculated the data of the generated figures.

These tools may be useful in any project.

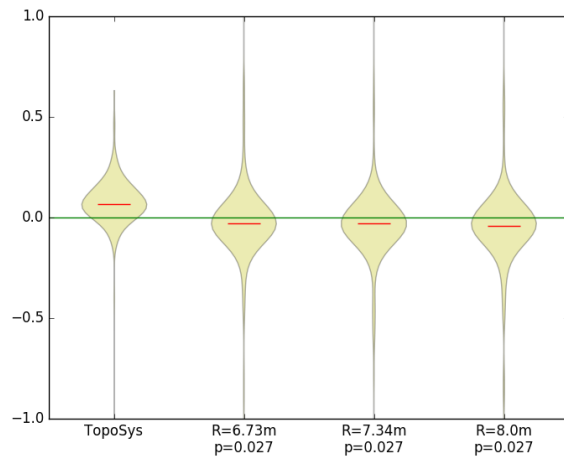


Figure 12. Another violin plots of the distribution of LiDAR-based surcaces compared to geodetic survey.

REFERENCES

- [1] G. Nagy, *A diagram to illustrate the distribution of slope and aspect of an area.* Óbudai Egyetem, 2016, pp. 24–27.
- [2] G. Van Rossum *et al.*, “Python Programming Language.” in *USENIX Annual Technical Conference*, vol. 41, 2007.
- [3] K. J. Millman and M. Aivazis, “Python for scientists and engineers,” *Computing in Science & Engineering*, vol. 13, no. 2, pp. 9–12, 2011.
- [4] T. E. Oliphant, “Python for scientific computing,” *Computing in Science & Engineering*, vol. 9, no. 3, 2007.
- [5] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, May-Jun 2007.
- [6] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, “The numpy array: a structure for efficient numerical computation,” *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.
- [7] F. Warmerdam, “The geospatial data abstraction library,” *Open source approaches in spatial data handling*, pp. 87–104, 2008.
- [8] J. L. Hintze and R. D. Nelson, “Violin plots: a box plot-density trace synergism,” *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

Integrity analysis of the RTCA tropospheric delay model

Sz. Rozsa, B. Ambrus and I. Juni

Department of Geodesy and Surveying

Budapest University of Technology and Economics, Faculty of Civil Engineering
Budapest, Hungary

rozsa.szabolcs@epito.bme.hu, ambrus.bence@epito.bme.hu, juni.ildiko@epito.bme.hu

Abstract—Electromagnetic signals broadcast by GNSS satellites suffer considerable delays while travelling through the atmosphere. Apart from the ionosphere, the troposphere also has a significant effect on the propagation. The delay caused can be separated into two different parts: the effect of gases in hydrostatic equilibrium and the effect of water vapour and condensed water present in the troposphere.

In navigation applications of GNSS not only the accuracy of the positioning needs to be known, but the integrity of the positioning service should be evaluated, too. The integrity information includes the maximum positioning error at an extremely rare probability level (10^{-7}), called the protection level. The RTCA (Radio Technical Commission for Aeronautics) specifies the minimum operational performance standard for GNSS systems used in the aeronautics. This standard recommends 0.12 m as the maximum tropospheric error in terms of standard deviation in the zenith direction, but it neglects both the geographical and seasonal variation of this error.

Our study focuses on the derivation of a new integrity model for the troposphere, which takes into consideration both the seasonal and geographical behaviour of the model performance using the extreme value theory.

The results show that the original RTCA recommendation is too conservative. Our study shows that the standard deviation is in the order of only 5 cm with a seasonal amplitude of 2-3 cm at the mid-latitudes. The application of the derived – more realistic – integrity model helps to improve the availability of GNSS positioning service in aviation.

I. INTRODUCTION

The global navigation satellite systems (GNSS) use range observations between the satellites and the receivers to derive the position of the user. These ranges are measured by measuring the duration of signal propagation and the results is multiplied by the velocity of light in vacuum to obtain the distance between the satellite and the receiver.

It is well known that radio waves propagate slower in the lower neutral part of the atmosphere, therefore this atmospheric layer (i.e. the troposphere) causes a significant signal delay. This delay is called tropospheric delay and it is modelled with models derived from various meteorological observations.

To assess the integrity of the satellite signal, the performance of these tropospheric delay models must be evaluated on an extremely rare probability level to ensure that the safety-of-life users (e.g. aviation,) can absolutely rely on the coordinates provided by the GNSS receivers.

Error models used in current ‘standard’ for safety-of-life GNSS [1] applications are considered very conservative when

it comes to residual error modelling. In recent times, there has been much interest in revisiting these models with the aim of making them less conservative in order to assess the availability of satellite positioning more reliably.

The current tropospheric delay model from the RTCA Satellite-Based Augmentation System (SBAS) Minimal Operations Standards (RTCA MOPS) [1] possesses an associated residual error that is equal to 0.12 meters in the vertical sense. The value is derived from the results reported in [2]. While this approach gives a resulting standard deviation that is much higher than the estimated standard variance that best fits the data (0.05 m), it can surely be considered conservative for most applications. [2] also states however, that characterizing the delay errors beyond the $\pm 4\sigma$ level using a normal distribution is not recommended as it drastically underestimates the true distribution. The probability level denoted by $\pm 4\sigma$ corresponds to 99.994% which is obviously high, however safety critical systems may demand even higher levels. These considerations leave room for doubt whether the current model is safe to use under all circumstances. Additionally, the current residual error model has also been inspected in [3], where it is concluded that the model seems to be too conservative. Furthermore, it also lacks the ability to take into account the latitude dependency of the tropospheric delay estimations.

In near future, more demanding applications are expected to arise and as most of these will be based on multi-frequency and multi-constellation use of GNSS, they suffer from ionospheric delays less than today. This creates a demand for more accurate tropospheric error modelling and ensures its importance in approximating integrity while maintaining sufficient system availability. Recent investigations have already been done on the performance of the European Geostationary Navigation Overlay Service (EGNOS) in aiding localizer performance and vertical guidance (LPV) approaches of airplanes [4]. The calculation and validation of the protection levels established using such an overlay service has also been of interest recently, using open-source software for the computation [5].

The approach proposed in this paper can be summarized as analyzing tropospheric delay data using state-of-the-art knowledge on tropospheric modelling, in order to characterize the performance of the RTCA MOPS model by simple overbounding models that safety-of-life users can employ to derive error bounds on their positioning performance (e.g. in the form of protection levels). To this end we employed a dedicated processing methodology using reference dataset generated by a raytracing algorithm on numerical weather models and a combination of statistical concepts and techniques to rigorously prove the correctness of error bounds to an associated confidence level. To establish the overbounding relation

between the model and the reference data and deal with the tails of the distribution, the extreme value theory was employed.

II. THE RTCA-MOPS TROPOSPHERIC DELAY MODEL

The tropospheric delay model described in [1] calculates the total slant delay for satellite i as:

$$TC_i = (d_{\text{hyd}} + d_{\text{wet}}) \cdot m(El_i), \quad (1)$$

where TC_i denotes the total tropospheric delay [m], d_{hyd} and d_{wet} correspond to the hydrostatic and wet part of the delay in the zenith direction [m], while $m(El_i)$ is the value of the mapping function [-] at a given El elevation angle that is used to scale the zenith delay to the actual elevation angle.

The hydrostatic and wet parts of the delay are computed from the receiver's height and the estimation of five meteorological parameters: air pressure, temperature, water vapour pressure, temperature lapse rate and water vapour lapse rate. Each parameter (ξ) is estimated for the receiver's latitude (ϕ) and day-of-year (DOY) from the mean value (ξ_0) and its seasonal variation ($\Delta\xi$):

$$\xi(\phi, D) = \xi_0(\phi) + \Delta\xi(\phi) \cdot \cos\left(\frac{2\pi(DOY - DOY_{\min})}{365.25}\right). \quad (2)$$

The value of DOY_{\min} is different for the northern and southern hemisphere. The model works with a predefined value set for each meteorological parameter given for latitudes 15° (or less), 30° , 45° , 60° and 75° (or greater) and linearly interpolates for intermediate latitudes using the two closest values. The equation of the mapping function used to scale the zenith delays to slant range is the same as equation (5) for the integrity calculation.

III. INTEGRITY MODELLING IN RTCA-MOPS

According to [1], the following formula is used in the RTCA-MOPS to calculate the residual error for GPS pseudorange measurements for satellites used for the positioning:

$$\sigma_i^2 = \sigma_{i,\text{flt}}^2 + \sigma_{i,\text{UIRE}}^2 + \sigma_{i,\text{air}}^2 + \sigma_{i,\text{tropo}}^2, \quad (3)$$

where:

- σ_i is the standard deviation of satellite i pseudorange measurement [m],
- $\sigma_{i,\text{flt}}^2$ is the model variance of the residual errors for fast and long-term corrections [m],
- $\sigma_{i,\text{UIRE}}^2$ is the model variance of the slant range ionospheric delay estimation error [m],
- $\sigma_{i,\text{air}}^2$ is variance of the airborne receiver errors [m],
- $\sigma_{i,\text{tropo}}^2$ is the variance of tropospheric delay estimation error [m].

The standard deviation of the residual tropospheric error is modeled as a random integer with the standard deviation of $\sigma_{i,\text{tropo}}$, which is calculated as:

$$\sigma_{i,\text{tropo}} = (\sigma_{\text{TVE}} \cdot m(\theta_i)), \quad (4)$$

$$m(\theta_i) = \frac{1.001}{\sqrt{0.002001 + \sin^2(\theta_i)}}, \quad (5)$$

where σ_{TVE} denotes the vertical residual error of the tropospheric delay estimation and is equal to 0.12 meters and θ_i is the satellite elevation angle. Note that the vertical residual error of the tropospheric delay estimation is a constant value which globally overbounds the standard deviation of the residuals, but as it neglects the effect of latitude on the accuracy of the tropospheric delay estimation, leads to an overly conservative model in many regions.

Combining these terms, one ends up with the variance of the total residual error which enables the system to calculate the horizontal and vertical protection levels (HPL and VPL) for a given position as follows:

$$HPL = K_H \cdot d_{\text{major}}, \quad (6)$$

$$VPL = K_V \cdot d_{\text{major}}, \quad (7)$$

where K_H and K_V are constants depending on the different approach type and d_{major} [m] corresponds to the uncertainty along the semimajor axis of the error ellipse:

$$d_{\text{major}} \equiv \sqrt{\frac{d_{\text{east}}^2 + d_{\text{north}}^2}{2} + \sqrt{\left(\frac{d_{\text{east}}^2 - d_{\text{north}}^2}{2}\right)^2 + d_{\text{EN}}^2}}. \quad (8)$$

The terms in the equation stand for the following:

- d_{east}^2 is the variance of model distribution that overbounds the true error distribution in the east axis [m²],
- d_{north}^2 is the variance of model distribution that overbounds the true error distribution in the north axis [m²],
- d_{EN}^2 is the covariance of the model distribution in the east and the north axes [m²],
- d_U^2 is the variance of model distribution that overbounds the true error distribution in the vertical axis [m²].

All the model variances are calculated using the partial derivatives of the position error in the respective direction with respect to the pseudorange error on each satellite.

Using the HPL and the VPL values, the instrument can decide whether current accuracy of the position is suitable for navigational purposes during the different approach types.

IV. REFERENCE DATA

A. Meteorological data

In order to assess model performance, a reference data set of tropospheric delays was needed. Four European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim solutions per day were used to calculate this data set with ray-

tracing the various atmospheric layers. Relative humidity, temperature and geopotential values estimated on 37 pressure levels (from 1000 hPa to 1 hPa) with a resolution of $1^\circ \times 1^\circ$ were collected for the years 2000-2016 for this study. Besides ECMWF ERA-Interim solutions International Standard Atmosphere (ISA) [6] values were used to expand the atmospheric profiles up to the height of 86 km.

B. Computation of reference tropospheric delays

The ray-tracing method supposes specific layers of the atmosphere, where the path of a beam is traced. The beam starting at a certain elevation angle continuously refracts at different layers and changes direction [7]. The tropospheric delay can be calculated by multiplying the length of the refracted beam with the refractivity in the given layer.

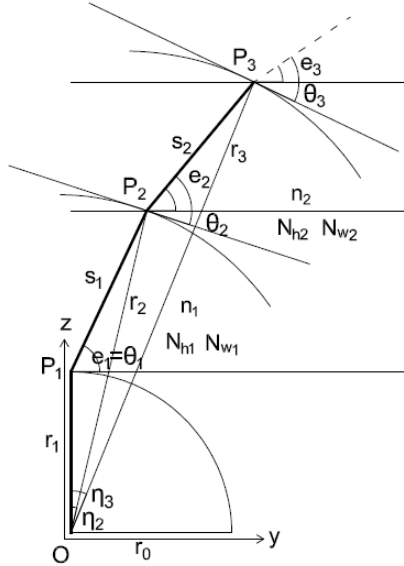


Fig. 1. The principle of the ray tracing showed with a beam starting at the surface of a sphere (modelling the Earth) and refracting at each layer of the atmosphere with different refractivity

To obtain optimal results, the resolution of the meteorological data needs to be increased. The interpolation is done linearly for the temperature and exponentially for the air pressure and water vapour pressure. Then the hydrostatic and the wet refractivity can be calculated for each layer as well as the distance travelled in the layer. The hydrostatic and wet delays are defined:

$$dS_{h,w} = \sum_{i=1}^{k-1} s_i \cdot N_{h,w,i}, \quad (9)$$

where s_i is the length of the refracted beam [m] and $N_{h,w,i}$ is the hydrostatic and wet refractions [-] in the i -th layer.

V. METHODOLOGY

A. Principles

The general integrity requirements of radio navigational aids used in civil aviation is formulated in [8]. According to this document, the integrity of GNSS positioning service must be evaluated at the extremely rare probability level of 2×10^{-7}

in any approach. Assuming the duration of an average approach of 150 seconds and no concurrent approaches in the same time, the recurrence interval of an integrity event would be 25 years.

Since only a limited number of observation samples are available to assess the performance of the tropospheric delay models, one must use a probabilistic approach for such a study. It would be straightforward to fit a normal distribution to the residuals of the estimated tropospheric delays, and extrapolate it to the tails of the distribution. However, the probability plot of the residuals (Fig. 2) clearly indicates that the tails of the residuals significantly deviate from the normal distribution. Thus, the extreme value theory must be applied for this problem.

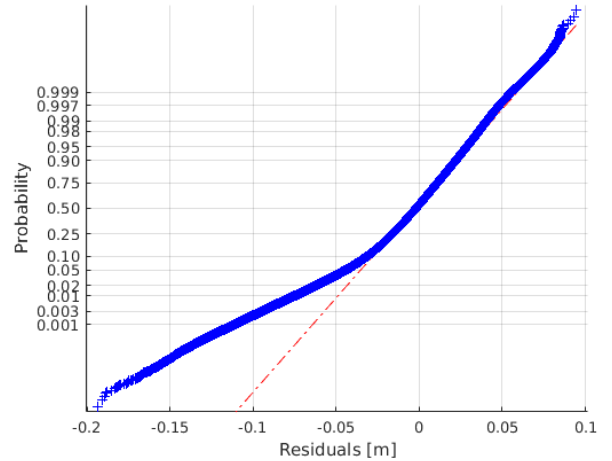


Fig. 2. Normal probability plot of the hydrostatic tropospheric delay model residuals for the latitude band N40°-N50°

B. Principles of Extreme value theory

The Fisher-Tippett theorem states that the maximum of a sample of independent and identically distributed probability variables after proper renormalization can converge to one of the three possible distributions, the Gumbel, the Fréchet or the Weibull distribution.

The three distribution functions are the following:

$$H(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \exp\{-x^{-\alpha}\} & \text{if } x > 0 \end{cases}, \quad (10)$$

for the Fréchet,

$$H(x) = \begin{cases} \exp\{-(-x)^{-\alpha}\} & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}, \quad (11)$$

for the Weibull, and

$$H(x) = \exp\{-\exp\{-x\}\} \quad x \in R, \quad (12)$$

for the Gumbel distribution.

The general extreme value (GEV) theory [9] combines the previous three distributions to the general extreme value distribution. The distribution function is:

$$H(x) = \begin{cases} \exp\{-[1 - k(x - \xi)/\alpha]^{1/k}\} & \text{if } k \neq 0 \\ \exp\{-\exp\{-(x - \xi)/\alpha\}\} & \text{if } k = 0 \end{cases} \quad (13)$$

with x bounded by $\xi + \alpha/k$ from above if $k > 0$ and from below if $k < 0$. Here ξ and α are the location and scale parameters, while k is the shape parameter. The shape parameter determines which original extreme value is represented by the GEV distribution:

- for $k > 0$ the Fréchet distribution (heavy tailed)
- for $k = 0$ the Gumbel distribution (light tailed)
- for $k < 0$ the short tailed negative Weibull distribution

is described by the GEV distribution.

C. Estimation of extreme tropospheric error using GEV theory

To study the performance of tropospheric delay models under extreme conditions, firstly, the tropospheric model error must be calculated. To achieve this, the hydrostatic and wet tropospheric delays were computed using the RTCA-MOPS troposphere model based on surface meteorological parameters obtained from the numerical weather models. Since numerical weather model data are given in constant pressure levels instead of elevation levels, therefore an interpolation or extrapolation of the air pressure, water vapour pressure and the ambient temperature was needed to calculate the parameters on the ground.

Afterwards these tropospheric delays were subtracted from the reference values calculated with ray-tracing the entire atmosphere. These residuals were calculated in 18, equally sized latitude bands for the whole globe. Fig. 3. shows the time series of the hydrostatic delay residuals for all the grid points in the latitude band between N41 to N50 latitudes. The figure shows, that both the spread of the daily residuals have a significant seasonal variation. To derive an appropriate model for the integrity assessment, this seasonal variation must be removed from the residuals and later restored in the derived model to be able to represent the seasonal behavior

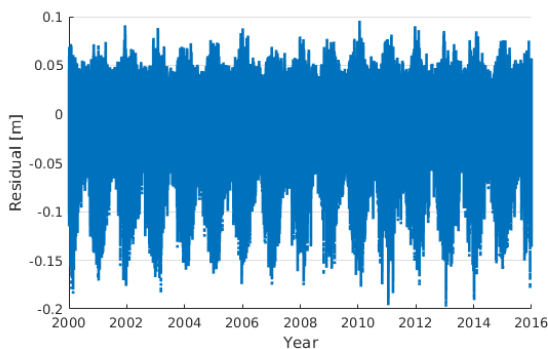


Fig. 3. Time series of the residuals of the hydrostatic delays with w.r.t. the raytraced reference values

of the tropospheric model performance. Basically, this is equivalent with the normalization of the time series of the residuals. Thus, the daily standard deviation of the residuals was calculated and a periodic function was fit to these mean and standard deviation values considering both the annual and the semi-annual components of the seasonal variations (Fig. 4).

The model function for the daily standard deviation values:

$$\begin{aligned} \sigma(DOY) = & \bar{\sigma} + A_1 \cos\left(\frac{DOY - DOY_0}{365.25} 2\pi\right) + \\ & + A_2 \sin\left(\frac{DOY - DOY_0}{365.25} 2\pi\right) + \\ & + A_3 \sin\left(\frac{DOY - DOY_0}{365.25} 4\pi\right) + \\ & + A_4 \sin\left(\frac{DOY - DOY_0}{365.25} 4\pi\right), \end{aligned} \quad (14)$$

where the unknown parameters are: $\bar{\sigma}$ is the mean value of the daily mean residuals for the total time series, DOY_0 is the day of the annual minimum of the standard deviation of the daily residuals (the phase), while A_1 and A_2 are the amplitudes of annual, and A_3 and A_4 of the semi-annual terms of the seasonal variation.

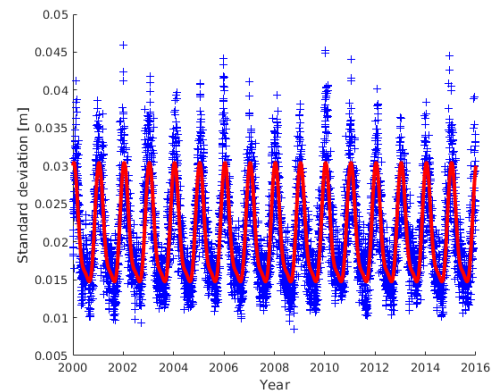


Fig. 4. The seasonal variation of the daily standard deviations of the residuals and the fitted model

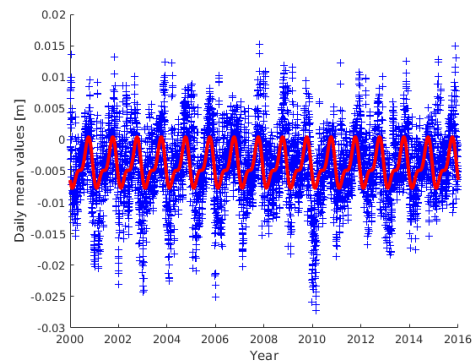


Fig. 5. The seasonal variation of the daily mean values of the residuals and the fitted model

Afterwards, the residuals (δ) were normalized using a zero-mean assumption with the following equation:

$$\delta_n = \frac{\delta}{\sigma(DOY)}. \quad (15)$$

In the next step, the normalized residuals were used for extreme value analysis. Since the samples covered 17 years of data, 17 annual extremes (maximum and minimum values)

were identified and selected for the extreme value analysis. The GEV distribution was fit to these extremes using the MATLAB software [8], and finally the extreme value representing the recurrence time of 25 years was estimated using the fitted distribution for both the maximal (positive) and minimal (negative) extremes. From these two values, the one with the larger absolute value was chosen as the maximal expected error of the normalized residuals ($\Delta_{n, \max}$).

Since the RTCA-MOPS proposes a calculation of the protection levels based on the standard deviation of parameters defined as normally distributed probabilistic variables, the previously estimated extreme values had to be converted to the standard deviation of normally distributed probabilistic variables. Thus:

$$\sigma_{n, \max} = \frac{\Delta_{n, \max}}{K} \quad (16)$$

where K is the value of the probability density function of the standard normal distribution at the probability level (meaning the probability of non-exceedance) of $1-10^{-7}$.

To estimate the seasonal variations of the troposphere model errors the following overbounding model is formulated for each latitude band:

$$\sigma_{\max}(DOY, band) = \frac{\Delta_0}{K} + \sigma(DOY) \cdot \sigma_{n, \max}, \quad (17)$$

where Δ_0 is an offset parameter, that is necessary for achieving the overbounding of model error. This offset parameter is calculated by fitting another extreme value distribution function to the annual extremes of the daily mean values (Fig. 5).

VI. RESULTS

The overbounding models of the troposphere model error were calculated for all the 18 latitude bands of the global grid for both the hydrostatic and the wet component of the tropospheric delay models (TABLE I. and TABLE II.). The results can be seen on Fig. 6 for the northern hemisphere for both components

The figures indicate that the σ_{\max} values of the hydrostatic and wet components are in all scenarios better than ± 6 and ± 10 centimeters, respectively. Since the total delay can be computed as the sum of the two components, the maximum standard deviation of the total tropospheric delay error can be computed using the law of error propagation:

$$\sigma_{TD, \max} = \sqrt{\sigma_{HD, \max}^2 + \sigma_{WD, \max}^2} \cong 0.12 \text{ m} \quad (18)$$

where $\sigma_{HD, \max}$ and $\sigma_{WD, \max}$ are the maximum standard deviation for the hydrostatic and wet delay in the zenith direction, respectively. This value perfectly agrees with the recommendations of the RTCA-MOPS. However, Fig. 6. shows that this value is too conservative for large regions of the world.

The maximum tropospheric error, i.e. the integrity model for the tropospheric delays, can be estimated by reformatting Eq. (17):

$$\Delta_{\max}(DOY, band) = \Delta_0 + \sigma(DOY) \cdot K \cdot \sigma_{n, \max}. \quad (19)$$

TABLE I.
INTEGRITY MODEL PARAMETERS FOR THE HYDROSTATIC DELAY

Model parameters								
Band	Δ_0 [mm]	$\bar{\sigma}$ [mm]	A_1 [mm]	A_2 [mm]	A_3 [mm]	A_4 [mm]	DOY_0 [day]	$\sigma_{n, \max}$
<i>Northern hemisphere</i>								
90–81	87.8	14.1	2.8	0.4	-0.2	0.2	2	2.0
80–71	51.0	21.6	6.0	1.6	-0.1	0.4	0	1.3
70–61	43.2	22.9	8.4	1.5	0.1	0.0	0	1.3
60–51	29.7	24.3	10.0	1.8	0.5	0.1	1	1.5
50–41	26.6	20.9	7.0	2.5	2.0	0.7	0	1.7
40–31	20.7	15.6	1.3	1.8	2.3	1.1	0	2.1
30–21	15.2	11.6	-3.6	0.4	1.5	1.0	3	2.7
20–11	16.0	7.1	-2.1	0.1	0.6	0.4	8	3.9
10–0	17.5	4.6	-0.2	-0.1	0.4	0.2	1	3.3
<i>Southern hemisphere</i>								
1–10	17.3	5.0	-0.2	-0.5	0.4	0.2	3	2.6
11–20	15.3	6.7	0.8	-0.3	0.5	0.4	2	3.6
21–30	10.6	10.2	0.3	-0.9	0.7	0.5	2	2.3
31–40	21.1	16.4	-2.8	-1.6	0.5	0.1	0	2.0
41–50	41.8	25.1	-3.4	-1.5	0.0	0.0	0	1.4
51–60	73.9	31.3	-3.4	-1.3	-0.9	0.4	2	1.3
61–70	101.1	26.6	-5.2	-2.1	-1.0	0.5	0	1.8
71–80	97.1	23.0	-8.6	-5.4	-0.3	-0.4	1	2.8
81–90	92.4	13.2	-5.4	-3.3	-0.3	0.0	1	4.0

TABLE II.
INTEGRITY MODEL PARAMETERS FOR THE WET DELAY

Model parameters								
Band	Δ_0 [mm]	$\bar{\sigma}$ [mm]	A_1 [mm]	A_2 [mm]	A_3 [mm]	A_4 [mm]	DOY_0 [day]	$\sigma_{n, \max}$
<i>Northern hemisphere</i>								
90–81	70.4	8.5	-3.8	-2.7	0.8	1.5	6	2.9
80–71	54.6	15.5	-5.6	-3.5	1.1	1.5	1	1.9
70–61	55.7	22.3	-6.7	-3.9	1.8	1.5	2	1.6
60–51	59.8	29.0	-6.0	-4.5	1.8	1.4	3	1.2
50–41	60.2	37.3	-6.1	-5.8	0.8	1.2	1	1.1
40–31	72.5	47.7	-10.7	-6.7	2.1	1.1	2	1.0
30–21	89.9	59.7	-13.6	-5.1	2.8	0.0	0	0.8
20–11	117.6	57.0	-1.2	-1.4	1.3	-5.4	0	1.0
10–0	58.6	46.8	6.7	1.6	1.1	2.9	1	0.9
<i>Southern hemisphere</i>								
1–10	74.6	55.3	2.4	-6.5	3.4	-2.0	2	0.7
11–20	120.1	61.0	9.0	2.2	2.0	-1.3	1	0.9
21–30	100.8	53.6	9.5	3.9	1.3	1.0	0	0.8
31–40	111.3	42.6	7.0	5.1	0.1	1.1	2	0.9
41–50	97.1	34.1	4.6	4.5	-0.2	0.7	0	1.1
51–60	94.6	25.1	2.3	3.0	-0.5	0.5	1	1.1
61–70	86.4	17.2	1.0	1.5	-0.4	0.2	2	1.3
71–80	60.8	13.9	6.6	4.4	-0.8	-0.2	1	2.5
81–90	48.2	9.2	5.9	3.8	-0.7	-0.5	3	5.1

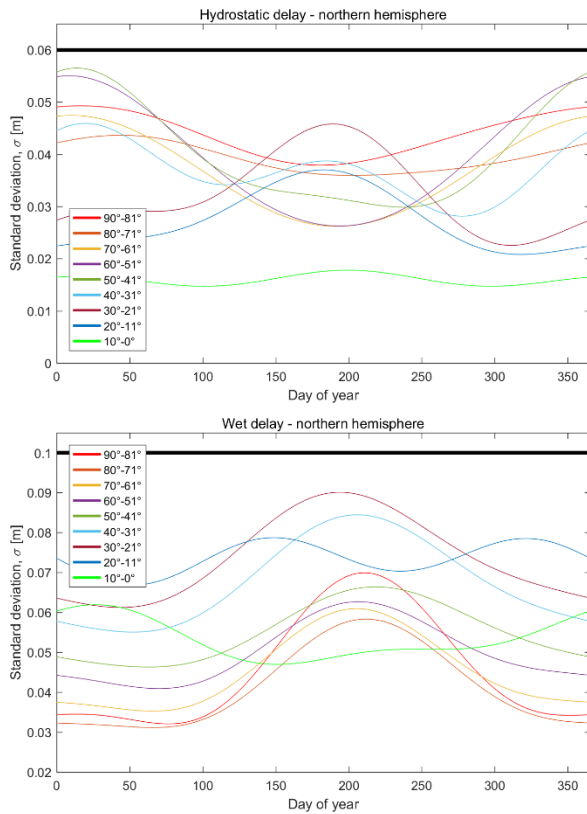


Fig. 6. The seasonal variation of the σ_{\max} in the latitude bands on the northern hemisphere

Fig. 7 depicts the unnormalized hydrostatic and wet residuals and the derived integrity model for the latitude band between N40° and N50° latitudes. It can be clearly seen that the derived model truly overbounds the tropospheric delay error and it is significantly less conservative than the original RTCA model. Moreover, the derived model takes into consideration the seasonal variations of the tropospheric delays caused by the climate.

VII. CONCLUSION

The results of our study confirmed that the RTCA MOPS recommendation of 0.12m for modelling the maximal tropospheric delay error in the zenith direction is appropriate, but it can also be stated that it is too conservative for a large part of the globe.

In this paper, a less conservative, nevertheless reliable model was derived for the globe, which provides the users a more realistic limit for the maximal error of the tropospheric models. This leads to smaller level of the expected error, thus a smaller protection level for the assessment of the integrity of the system, which increases the availability of the satellite positioning services for safety-of-life users.

ACKNOWLEDGMENT

The authors thank the support of the European Space Agency and the Hungarian Space Office provided in the frame of the INTRO (INtegrity of TROposphere models) project funded by the ESA contract 4000114534/15/NL/NDe.

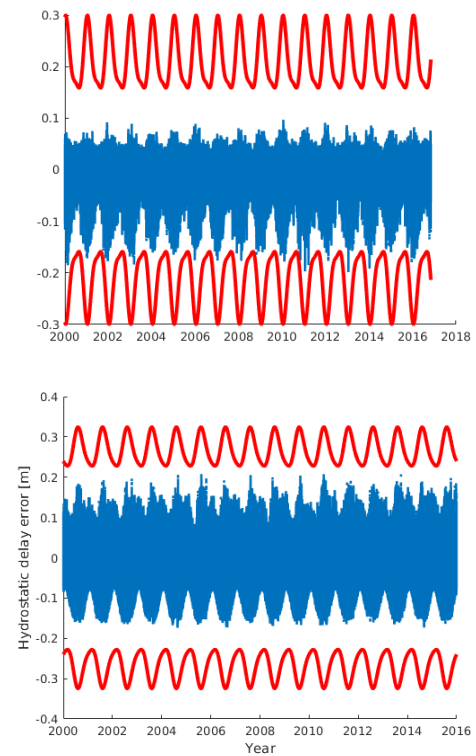


Fig. 7. The integrity model of the hydrostatic (top) and the wet (bottom) tropospheric delays for the latitude band N40°-N50°.

REFERENCES

- [1] *Minimum Operational Performance Standards for Global Positioning System/Satellite-Based Augmentation System Airborne Equipment*, RTCA DO-229, 2006.
- [2] J. P. Collins, R. B. Langley, "The residual tropospheric propagation delay: How bad can it get?", *11th International Technical Meeting of the Institute of Navigation*, Nashville, Tennessee, 1998.
- [3] S. Storm van Leeuwen, H. van der Marel, M. Toussaint, A. Martelluci, "Validation of SBAS MOPS troposphere model over the EGNOS service area", *European Navigation Conference (GNSS-2004)*, Rotterdam, The Netherlands, 2004.
- [4] R. Markovits-Somogyi, B. Takács, A. de la Fuente, P. Lubrani, "Introducing E-GNSS navigation in the Hungarian Airspace – The BEYOND experience and the relevance of GNSS monitoring and vulnerabilities", in *Selected papers of 3rd International Conference on Research, Technology and Education of Space*, Budapest, Hungary, 2017.
- [5] B. Takács, Z. Siki, R. Markovits-Somogyi, "Extension of RTKLIB for the calculation and validation of protection levels", in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-4/W2, 2017.
- [6] *Standard Atmosphere*, ISO 2533:1975, 1975.
- [7] J. Boehm, H. Schuh, "Vienna Mapping Functions", in *16th Working Meeting on European VLBI for Geodesy and Astrometry*, pp. 131-143, 2003.
- [8] International Civil Aviation Organization (ICAO), *Aeronautical Telecommunication*, vol. 1., Radio Navigation Aids, in *Annex 10 to the Convention on International Civil Aviation*, p.578, 2006
- [9] Jenkinson, A.F, The frequency distribution of the annual maximum (or minimum) of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, Vol 81., pp 158-171., 1955
- [10] MATLAB Statistics and Machine Learning Toolbox Release 2016a, The MathWorks Inc., Natick, Massachusetts, United States.

Modeling informant agreement in adolescents and parents: the example of callous-unemotional traits

Csilla Bozsik¹, Julia Gadoros², Judit-Inantsy Pap¹, Peter Vida³ and Jozsef Halasz^{2,4}

1. University of Debrecen, Institute of Psychology, Debrecen, Hungary

2. Vadaskert Child Psychiatry Hospital, Budapest, Hungary

3. Semmelweis University, School of Ph.D. Studies, Budapest, Hungary

4. Obuda University, Alba Regia Technical Faculty, Szekesfehervar, Hungary

bozsikcsilla@gmail.com

gadorosju@gmail.com

inantsy-pap.judit@arts.unideb.hu

fitz026@yahoo.co.uk

halasz.jozsef@amk.uni-obuda.hu

Abstract— Introduction. According to literature data, informant agreement might have major importance in the interpretation of different psychological states. In the present paper, a model of informant agreement was established on callous unemotional traits in non-clinical adolescents.

Methods. The Inventory of Callous-Unemotional Traits (ICU) was assessed in 279 primary and secondary school adolescents (age range: 11-16 years; 14.2±1.5, mean±SD; girls: 132) after informed consent. Additional to socioeconomic measures, both the self-report and parent versions of the ICU was assessed. The study was part of the data collection series about externalization problems in adolescents, in a collaboration between University of Debrecen and Vadaskert Child Psychiatry Hospital.

Results. The Cronbach α of the Hungarian ICU self-report was similar to what has been observed in the parent report (0.768 vs. 0.818), and the internal structure was similar to what has been described earlier in other national reports. In spite of the similarities, latent differences between the structure of the two report types made the direct comparison of the three factors (callous, uncaring, unemotional) from two sources somewhat vulnerable. Neither the age, nor socioeconomic background was associated with the self-report scores, but a major gender effect was observed (girls scored less, $p<0.001$). Overall parent-report scores were significantly higher compared to self-report scores, irrespective to gender ($p<0.001$).

Conclusions. Results of Hungarian Inventory of Callous-Unemotional Traits analyses were almost identical with the results of foreign ICU analyses. Our data suggest that informant agreement in that particular case should focus on overall scores.

Keywords: Adolescent, Callous, Informant agreement, Inventory of Callous-Unemotional Traits, Uncaring, Unemotional.

I. INTRODUCTION

A. Informant agreement and Callous/Unemotional Traits in adolescents

According to literature data, informant agreement might have major importance on the evaluation and further clinical treatment of patients [1,2]. This effect have marked consequences in the evaluation of mental health condition in the case of children and adolescents. In recent years, multiple informant sources (parent, teacher and self-report) were used to evaluate further treatment strategies in children and adolescents with clinical needs. E.g., the widely used Strengths and Difficulties Questionnaire served as a model for determining internalizing and externalizing symptoms both in clinical and non-clinical samples, but at present the usage of both the self-report and parent report versions seems indispensable. While parent scores were consistently higher in externalizing (conduct and hyperactivity scales) domains, higher emotional and peer problems scores were observed in the self-report versions; moreover, the clinical features were also differentially correlated [3,4].

A specific case also might be observed in informant agreement, where a consistent bias from both types of informants can be observed. Albeit the description of psychopathic traits has a long research line, the description of callous-unemotional traits in its present format was described by Frick and coworkers in recent years [5]. Literature data described a major importance of callous-unemotional traits in children and adolescents with conduct disorder, and the presence of trait has a marked effect in later antisocial development. Callous-unemotional traits are characterized by lack of remorse and guilt, shallow or superficial expression of emotions, a lack of concern for the feelings of others, and a lack of concern regarding performance in important activities [6,7]. Thus, additional to the behavioral pattern, marked alterations of the affective and cognitive domains can be observed in affected children and adolescents. The importance of the

trait was also outlined with the inclusion of callous-unemotional traits as specifier for conduct disorder into the DSM-5 (the current American diagnostic system) [8,9].

The presence of the Callous-unemotional traits has a robust impact on the relations of the adolescents. The consequences of the behavior of youth with conduct disorder and callous unemotional traits might cause behavioral and affective changes within the responder (parent, teacher, peers), and the problematic interpretation of the feedback signals (from parents, teachers and peers) create a specific double bias for the clinician [10]. In clinical terms, the value of informant agreement is mainly the proper outline of children and adolescents with clinical needs. Interestingly, in contrast with the previously mentioned Strengths and Difficulties Questionnaire, to our best knowledge, no comparable general cut off points were outlined within the Inventory of Callous-Unemotional Traits, in relation with informant agreement.

B. Aims

The aim of the present study was to delineate a model for informant agreement on callous-unemotional traits in non-clinical adolescents. Both parent and self-report information were assessed in children and adolescents with low-average risk for later antisocial behavior. Both responder and gender effects were expected.

II. METHODS

The study was approved by the Unified Psychological Ethical Committee (EPKEB). The sample consisted 279 (girls: 132, boys: 147) Caucasian non-clinical primary and secondary school students between the age of 11 and 16 years (14.2 ± 1.5 years, mean \pm SD). The directors of the schools (primary and secondary schools from the County Borsod-Abaúj-Zemplén) were informed about the details of the study, then the parents were informed during the regular parent-teacher meetings. After informed consent of the students and the parents (98% of the parents agreed with the participation), socioeconomic data and the Hungarian version of the Inventory of Callous-Unemotional Traits (ICU) were assessed.

Both the self-report and parent report version of the questionnaire were used. The questionnaire contains 24 items, each of them can be evaluated by a Likert scale from 0 to 3 (0="not at all true"; 3="definitely true"). Across different languages and samples, the best fitting structure contains a general callous-unemotional core, and three factors (callousness, e.g.: "The feelings of others are unimportant to me"; unemotional, e.g.: "I hide my feelings from others"; uncaring, e.g.: "I try not to hurt the feelings of others" (inverted/reversed score item)). Altogether, the ICU contains 12 items with reversed scoring, and the original inventory was validated in several context [11-14]. The first preliminary data of the Hungarian ICU parent report was published in relation with subscales of

the Strengths and Difficulties Questionnaire in non-clinical adolescents, where a positive correlation pattern between parent reported callous-unemotional traits and behavioral problems were described [15]. In another study involving Hungarian non-clinical children and adolescents, ICU self-report scores were positively correlated with self-reported proactive aggression in Hungarian non-clinical adolescents [16].

Statistical analysis. Statistica 7.0 and SPSS 20.0 program packages were used to analyze datasets. Cronbach α values were evaluated for both self-report and parent report versions to determine the internal consistency. For the exploratory factor analysis, principal component analysis was used with varimax rotation, with eigenvalue=1.6. Maximum number of factors was set at $n=5$. General Linear Model was used to assess gender and type of responder differences. Where necessary, Newman-Keuls post hoc comparisons were also run. Spearman correlations were used to describe correlation pattern between self-report and parent report ICU versions. The level of significance was set at $p=0.05$.

III. RESULTS

The internal consistency was high in both self-report and parent report questionnaires, albeit the parent version had somewhat higher Cronbach α values (Table 1).

Within the Inventory of Callous-Unemotional Traits, both gender and type of responder differences were present (gender: $F_{(1,277)}=35.467$, $p<0.001$; type: $F_{(1,277)}=7.819$, $p<0.01$), but the interaction between gender and type was not significant. The 90 percent values (indication for clinical condition) were also determined, according to gender and responder type (Table 2). The effect of age and socioeconomic background was not significant.

The factor structure of the self-report and parent report ICU was similar, and three factors were delineated. Five items from the 24 items (Item 8, Item 10, Item 12, Item 21 and Item 24) had different subgroup position (Table 3 and Table 4). The exclusion of these factors did not significantly change factor weights.

The histograms of overall scores in gender split for self-report (Fig. 1), parent report (Fig. 2) and the comparison histogram for overall scores were also presented (Fig. 3).

Spearman correlations were also run between self-report and parent report overall scores. The correlation was higher in boys (Spearman $R=0.47$, $p<0.0001$) compared to girls (Spearman $R=0.34$, $p<0.0001$); and a positive correlation was also present in the whole population studied (Spearman $R=0.47$, $p<0.0001$). Neither the age, nor socioeconomic background was correlated significantly with the overall ICU scores.

Table 1. The internal consistency values (Cronbach α) of the Hungarian parent and self-report version of the Inventory of Callous/Unemotional Traits (ICU).

Population	Cronbach α	
	Parent	Self-report
Boys (n=147)	0.803	0.783
Girls (n=132)	0.798	0.709
Total (n=279)	0.818	0.768

Both the self-report and parent report showed a marked internal consistency in the presented population.

Table 2. Means and standard deviations (SD) of the Hungarian version of the ICU.

ICU total	Population	N	Mean	SD	Limit (90pc)
Self-report	Total	279	20.20#	7.2	
	Boys	147	21.97**	7.5	32
	Girls	132	18.23*	6.3	26
Parent	Total	279	21.59	8.1	
	Boys	147	23.97*	7.7	35
	Girls	132	18.93	7.7	29

Symbols indicate significant differences ($p < 0.05$). ICU, Inventory of Callous/Unemotional Traits; *, significantly different compared with girls; #, significantly different compared with parent report. Limit was calculated as 90 percentile value.

Table 3. Factor loadings, ICU self-report.

Items	Callousness	Uncaring	Unemotional
Item01*			0.50
Item02	0.33		
Item03*		0.74	
Item04	0.48		
Item05*		0.54	
Item06			0.47
Item07	0.52		
Item08*			0.65
Item09	0.52		
Item10	0.26		
Item11	0.45		
Item12	0.49		
Item13*		0.34	
Item14*			0.66
Item15*		0.70	
Item16*		0.49	
Item17*		0.43	
Item18	0.46		
Item19*			0.50
Item20	0.42		
Item21			0.59
Item22			
Item23*		0.72	
Item24*		0.36	

With the exception of Item08, the structure is identical to what was described earlier by Essau et al (2006). In the case of Item08, a switch between Callousness and Unemotional factors occurred. In later studies, Item02 and Item10 were removed from the ICU (Kimonis et al, 2008; Ciucci et al, 2014). *, Inverted items.

Table 4. Factor loadings, ICU parent report.

Items	Callousness	Uncaring	Unemotional
Item01*			0.62
Item02	0.42		
Item03*		0.80	
Item04	0.59		
Item05*		0.52	
Item06			0.67
Item07	0.41		
Item08*		0.48	
Item09	0.69		
Item10			0.16
Item11	0.63		
Item12			0.48
Item13*		0.32	
Item14*			0.60
Item15*		0.83	
Item16*		0.51	
Item17*		0.52	
Item18	0.44		
Item19*			0.30
Item20	0.58		
Item21	0.50		
Item22			0.72
Item23*		0.79	
Item24*			0.12

With the exception of 3 items, the structure is identical to what was described earlier by Essau et al (2006) (Item08, Callous/Uncaring switch, Item12, Callous/Unemotional switch, Item24, Uncaring/Unemotional switch). Compared with the Hungarian ICU self-report, the above 3 items showed differences. In later studies, Item02 and Item10 were removed from the ICU (Kimonis et al, 2008; Ciucci et al, 2014). *, Inverted items.

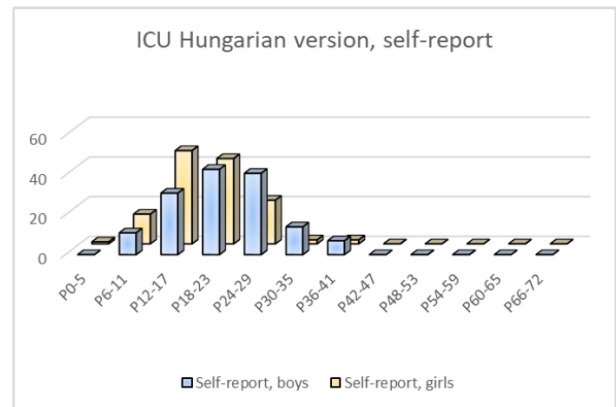


Fig. 1. Score histogram of the Hungarian version of the Inventory of Callous/Unemotional Traits, self-report. The theoretical maximum was 72 points. Axis X, score region values; axis Y, number of children.

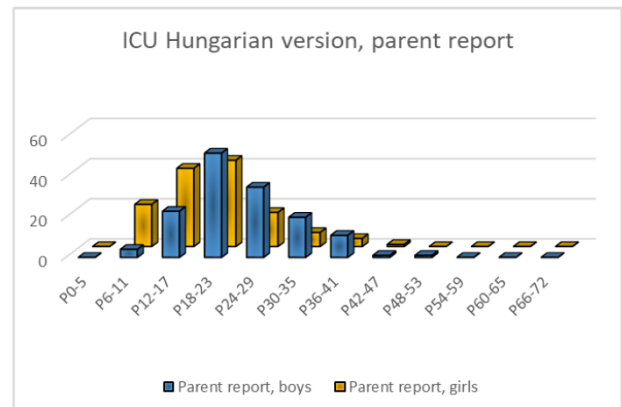


Fig. 2. Score histogram of the Hungarian version of the Inventory of Callous/Unemotional Traits, parent report. The theoretical maximum was 72 points. Axis X, score region values; axis Y, number of children.

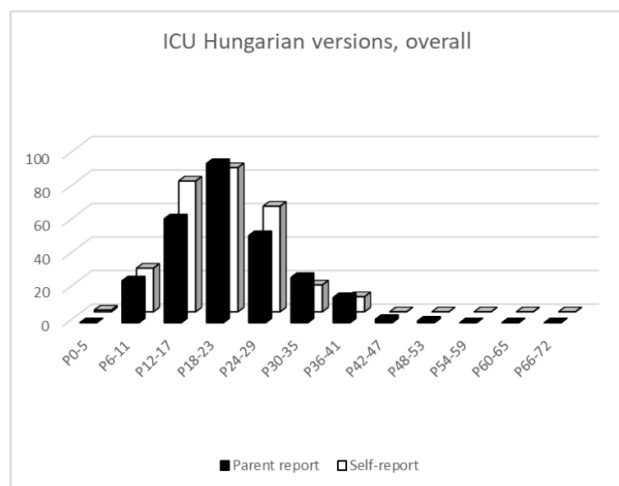


Fig. 3. Score histogram of the Hungarian version of the Inventory of Callous/Unemotional Traits, overall. The theoretical maximum was 72 points. Axis X, score region values; axis Y, number of children.

IV. DISCUSSION

The main results of the present study were the followings. First, in the present non-clinical population, similar factor structure of self-report and parent report ICU were observed. Second, parent report scores were significantly higher compared to self-report scores, and the scores of boys were significantly higher compared to girls, irrespective to the type of responder.

Callous-unemotional traits have major importance in the development of children with conduct disorder. Conduct disorder is not only more frequent in boys, but criminal involvement of boys with conduct disorder is also more prevalent [17-18]. Previous reports suggested a gender bias in incarcerated subjects, and a gender bias was also present in our study. Interestingly, the gender bias was observed in both the self-report and parent report results [10,12,14]. Most importantly, a more prominent correlation between the self-report and parent report scores were observed in boys.

Adolescents with conduct disorder and callous-unemotional traits are particularly vulnerable to later antisocial personality disorder. Unfortunately, no “A” treatment evidence is present in the case of antisocial personality disorder, thus the prevention of the condition has substantial importance [18]. Intervention based strategies in childhood and adolescence have major importance, and outline the relevance studying informant agreement in this particular condition.

The present non-clinical population can be considered as low-average risk study group in relation to antisocial development and considerably higher scores of callous-unemotional traits were described in incarcerated subjects or in a clinical population of adolescents with externalizing problems (e.g., compared to [12,14]). The present data were in line with the data of adolescents in low-average risk study groups [11].

The limitations of the study were the followings. First, the present study included only cross-sectional data, and

the age-effect of informant agreement within the sample could have been analyzed by longitudinal data collection. Second, the lack of detailed analysis of symptoms and conditions via structured diagnostic interview excluded the possibility to connect informant agreement to specific latent or subthreshold alterations. To our best knowledge, subthreshold psychopathologies were not assessed so far in relation with callous-unemotional traits, not even in cross-sectional settings. Third, the present study addressed adolescents with low-average risk, and a direct comparison with incarcerated or clinical subjects could deliver important additional data. In future studies, these issues also should be addressed.

V. SUMMARY

In the present paper, the informant agreement on the self-report and parent report versions of the Hungarian ICU were assessed. The informant agreement have a major role in this particular question, as agreement might be vulnerable from both sides: while the parents might serve as targets, the adolescents might have only a partial knowledge and insight of their own behavior. This behavioral pattern has crucial importance of later antisocial development, and the “double vulnerability” (parent bias and adolescent bias) might have crucial importance in our understanding and the wise interpretation of the symptoms.

VI. ACKNOWLEDGMENT

The authors declare no conflict of interest. The authors would like to express their gratitude to those students and families who participated in the project, and to J.P. Frick for providing freely the Inventory of Callous-Unemotional Traits.

REFERENCES

- [1] S. Vaz, R. Cordier, M. Boyes, R. Parsons, A. Joosten, M. Ciccarelli, M. Falkmer and T. Falkmer, “Is using the Strengths and Difficulties Questionnaire in a community sample the optimal way to assess mental health functioning?”, *PLoS One*, doi 10.1371-0144039, 2016.
- [2] L.L. Stone, R. Otten, R.C. Engels, A.A. Vermulst and J.M. Janssens, “Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: a review”, *Clin Child Fam Psychol Rev*, vol. 13, pp. 254-274, 2010.
- [3] K. Cleridou, P. Patalay and P. Martin, “Does parent-child agreement vary based on presenting problems? Results from a UK clinical sample”, *Child Adolesc Psychiatry Ment Health*, doi 10.1186-13034, 2017.
- [4] B. van Roy, B. Groholt, S. Heyerdahl and J. Clench-Aas, “Understanding discrepancies in parent-child reporting of emotional and behavioural problems: Effects of relational and socio-demographic factors”, *BMC Psychiatry*, doi 10.1186-1471-244, 2010.
- [5] P.J. Frick, A.H. Cornell, C.T. Barry, S.D. Bodin and H.A. Dane, “Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency”, *J Abnorm Child Psychol*, vol. 31, pp. 457-470, 2003.

- [6] P.J. Frick and S.F. White, "The importance of callous-unemotional traits in the development of aggressive and antisocial behavior", *J Child Psychol Psychiatry*, vol. 49, pp. 359–375, 2008.
- [7] P.J. Frick, J.V. Ray, L.C. Thornton and R.E. Kahn, "Annual Research Review: A developmental psychopathology approach to understanding callous-unemotional traits in children and adolescents with serious conduct problems", *J Child Psychol Psychiatry*, vol. 55, pp. 532–548, 2014.
- [8] American Psychiatric Association, "Diagnostic and Statistic Manual of Mental Disorders, DSM-5", New York, American Psychiatric Publishing, 2013.
- [9] E.R. Kimonis, K.A. Fanti, P.J. Frick, T.E. Moffitt, C. Essau, P. Bijttebier and M.A. Marsee, "Using self-reported callous-unemotional traits to cross-nationally assess the DSM-5 'with limited prosocial emotions' specifier", *J Child Psychol Psychiatry*, vol. 56, pp. 1249–1261, 2015.
- [10] S.F. White, K.R. Cruise and P.J. Frick, "Differential correlates to self-report and parent-report of Callous-Unemotional Traits in a sample of juvenile sexual offenders", *Behav Sci Law*, vol. 27, pp. 910–928, 2009.
- [11] C.A. Essau, S. Sasagawa and P.J. Frick, "Callous-unemotional traits in a community sample of adolescents", *Assessment*, vol. 13, pp. 464–469, 2006.
- [12] E.R. Kimonis, P.J. Frick, J.L. Skeem, M.A. Marsee, K. Cruise, L.C. Munoz, K.J. Aucoin and A.S. Morris, "Assessing callous-unemotional traits in adolescent offenders: validation of the Inventory of Callous-Unemotional Traits", *Int J Law Psychiatry*, vol. 31, pp. 241–252, 2008.
- [13] E. Ciucci, A. Baroncelli, M. Franchi, F.N. Golmaryami and P.J. Frick, "The association between Callous-Unemotional Traits and behavioral and academic adjustment in children: Further validation of the Inventory of Callous-Unemotional Traits", *J Psychopathol Behav Assess*, vol. 36, pp. 189–200, 2014.
- [14] E.R. Kimonis, K. Fanti, A. Goldweber, M.A. Marsee, P.J. Frick and E. Cauffman, "Callous-unemotional traits in incarcerated adolescents", *Psychol Assess*, vol. 26, pp. 227–237, 2014.
- [15] N. Pataky, A. Kormendy, C. Bozsik, J. Inantsy-Pap, J. Halasz and J. Gadoros, "Investigation of callous/unemotional traits and interpersonal characteristics among Hungarian adolescents – preliminary research results", *Psychiatria Hung*, vol. 26, pp. 427–433, 2011.
- [16] C. Bozsik, A. Kormendi, J. Inantsy-Pap, N. Kormendi, J. Gadoros, J. Halasz, "The relationship between reactive/proactive aggression, callous/unemotional traits and behavioural problems in Hungarian adolescents", *Psychiatria Hung*, vol. 28, pp. 48–56, 2013.
- [17] P. Vida, J. Halasz and J. Gadoros, "Aggressive and prosocial behavior in childhood psychopathology", *Psychiatria Hung*, vol. 28, pp. 261–273, 2013.
- [18] National Institute for Health and Care Excellence, "Antisocial behaviour and conduct disorder in children and young people. Recognition, intervention and management", British Psychological Society and The Royal College of Psychiatrists, *National Clinical Guideline Number 158*, pp. 1–468, 2013.

MATLAB and Python in Teaching Calculus

Árpád Horváth

Alba Regia Technical Faculty of Óbuda University,

H-8000 Székesfehérvár, Budai út 45., Hungary

Email: horvath.arpad@amk.uni-obuda.hu

Abstract—This year Obuda University have been started a course for the students to teach calculus with the aid of computer. As of many students learn programming in Python on an other course there is two group of students who use Python for learning calculus and three groups of students who use MATLAB for it. In this article I summarize the advantages and disadvantages of the symbolic mathematical module of both programming language.

I. INTRODUCTION

In the new curriculum of Obuda University there is a course unit named Mathematics I. that is common for most of the BSc students of the university. In that curriculum there is lessons (2 hours/week), practical course (2 hours/week) and there is a laboratory practice (1 hours/week in average). This laboratory practice lasts two hours in every second weeks or – for other groups – three hours in every third weeks.

Those students who learns Mathematics I. at the Alba Regia Tehchnical Faculty of the Obuda University learns Python[1], [2] as the first programming languages too in an other course.

There are many differences between the Python and MATLAB languages, so the students can be confused. Just to mention some differences: MATLAB has a 1-based indexing while Python 0-based, Python uses square brackets for indexing, MATLAB the round brackets; Python uses the `return` keyword for the return values of a function while MATLAB has a radically different method; Python uses the operator `**` to calculate the power, while MATLAB uses the `^` operator.

These and the many other differences can be confusing for the students who have not written any program code before. In this article I will discuss how easy they are to install at home and to learn.

It is worth to mention that using Python in the curriculum of the Obuda University is not new. There were a successful experiment on the Obuda University in teaching combinatorics using Python.[3]

II. INSTALLATION

A. MATLAB and Symbolic math toolbox

If one want to use MATLAB. They need to buy MATLAB, or to have an e-mail address that belongs to the domain of the university that has a MATLAB licence for its students. In our university there is a separate process to get an e-mail address that ends with uni-obuda.hu. If someone wants to get a Microsoft Office 365 as students, he or she also need an e-mail like that. The student can get the e-mail address at the MS Office homepage of the University. This is not well documented on the webpage of the university, were there is

a documentation about the installation. This is a quite good documentation except that.

If someone have the proper e-mail address he or she must register on the webpage of Mathworks, the developer of MATLAB and the software can be downloaded or used online in a web browser. For teaching calculus we need the Symbolic math toolbox[4] which is quite simple to install.

B. Python and sympy package

If you want to use python, you can use more Python distributions and Integrated Development Environments (IDE). In this article I will focus on installation on Windows. On Linux it is easier to install Python, but most of the students use Windows.

I installed most distributions of Python on Windows recently. The most robust combination was the one can be downloaded from the python.org with the PyCharm IDE to ease the usage. This Python distribution includes a simple IDE called IDLE. This includes an Interactive shell, like the Command Window in MATLAB, and one can edit Python files with it. It can fulfill the needs of the course. Without PyCharm you need the command line to install the sympy package, but it is not too difficult to do and a normal user without administrator privilege can do it. I created and shared a video to help the students to do it at home.

I think that PyCharm Professional is the most advanced IDE for Python right now. It can be reached for the students of universities without any fee for one years. For this one needs to have an e-mail address belonging to the domain of the university. The Professional version helps the advanced usage of Python, like web development (django, flask). But the most important features is available in the Community and Edu version of PyCharm. These versions are free for anyone and allows to install packages easily.

I have tried two other distributions on Windows: Anaconda and Enthought Python. The main goal of these packages is to help data analysis and big data applications. Both of them comes with a lot of packages by default including sympy. The installation of these distributions have difficulties on systems where there is a space or accentuated letter (á) in the path name. On one of my computers I had a user name with space and accent, and I have not managed to install that Python distributions there.

The screen shots were made by the IPython 6.1.0 interactive shell in Linux with Python 3.5.2 and SymPy 1.1.1. This is like the one can be found in the Anaconda Python distribution.

C. Sympy live

There is an online way to run sympy. In the live.sympy.org website we can do basic calculation with sympy. The mathematical expressions are rendered with $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ so they look like in a math book. However there is no possibility to run more complex code like cycles and function definition.

III. CREATING SYMBOLS

A. Symbolic math toolbox

Creating the real symbols x , y and t with the Symbolic math toolbox is as easy as:

```
syms x y t
```

We can add some assumptions for some special tasks:

```
syms a positive
syms l n integer
```

but we did not use this feature in the course.

B. SymPy package

When we use SymPy package we need to import its functions and the variables we need. If we want to render the expressions in a nice looking way (in the environment that support $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ or UTF-8) we can initialize the printing of expressions too:

```
from sympy import *
from sympy.abc import x, y, theta
init_printing()
```

IV. EXPANDING, SIMPLIFYING EXPRESSIONS, SUBSTITUTE VALUES, LIMITS OF THE FUNCTIONS

Both of the two examined tools (the Symbolic math toolbox and SymPy) has the same function (or method) names, but its usages is sometimes different. Both has the names `expand`, `simplify` and `subs`, `limit`. They have usually almost the same syntax. The main difference is that the `limit` is a method of an expression instance in Python as we can see in the example below.

They have some constants like `pi` and we can get the numerical values of expressions if we substituted all of the variables with a given value. If we want to give expressions the most important difference is that we use different operator for the power.

In the `limit` function we can use $\pm\infty$ as the value the variable goes to, but we have different sign for it. In MATLAB we use `Inf`, in sympy `oo` (two small O letters). If the right and left limit is different, MATLAB gives NaN value for the “simple” `limit` function. In that case Python uses right limit as default.

The examples for simplifying function and limit calculation can be found in the Figures 1 and 2, examples for expansion and substitution can be found in the Figures 3 and 4, for MATLAB and Python.

```
>> f = (x^2 - x - 6) / (2*x^2 - 18);
>> simplify(f)

ans =

(x + 2)/(2*(x + 3))

>> limit(f, x, -3)

ans =

NaN

>> limit(f, x, -3, 'right')

ans =

-Inf

>> limit(f, x, -3, 'left')

ans =

Inf
```

Figure 1. Simplifying function and limits in MATLAB

```
In [6]: f = (x**2 - x - 6) / (2*x**2 - 18)
In [7]: simplify(f)
Out[7]:
      x + 2
-----
2*(x + 3)

In [8]: limit(f, x, -3)
Out[8]: -∞

In [9]: limit(f, x, -3, "+")
Out[9]: -∞

In [10]: limit(f, x, -3, "-")
Out[10]: ∞

In [11]: limit(f, x, oo)
Out[11]: 1/2
```

Figure 2. Simplifying function and limits in Python

```
>> expand((x+y)^5)
ans =
x^5 + 5*x^4*y + 10*x^3*y^2 + 10*x^2*y^3 + 5*x*y^4 + y^5
>> subs(expand((x+y)^5), y, x)
ans =
32*x^5
```

Figure 3. Extension and substitution in MATLAB

```
In [41]: expand((x+y)**5)
Out[41]:
5      4      3 2      2 3      4      5
x  + 5·x ·y + 10·x ·y + 10·x ·y + 5·x·y + y
In [42]: expand((x+y)**5).subs(y, x)
Out[42]:
5
32·x
```

Figure 4. Extension and substitution in Python

```
>> f = x^sym(3/4);
>> diff(f)
ans =
3/(4*x^(1/4))
>> int(f)
ans =
(4*x^(7/4))/7
>> int(f, 1, 5)
ans =
(20*5^(3/4))/7 - 4/7
>> format long
>> double(int(f, 1, 5))
ans =
8.982004356806028
```

Figure 5. Derivative and integral of a function in sympy

```
In [129]: f = x**(Rational(3/4))
```

```
In [130]: f.diff()
```

```
Out[130]:
3
-----
4·√x
```

```
In [131]: f.integrate()
```

```
Out[131]:
7/4
-----
4·x
7
```

```
In [132]: f.integrate((x, 1, 5))
```

```
Out[132]:
3/4
- 4/7 + 20·5/7
```

```
In [133]: f.integrate((x, 1, 5)).n(26)
```

```
Out[133]: 8.9820043568060289143319010
```

Figure 6. Derivative and integral of a function in sympy

V. DIFFERENTIATION AND INTEGRATION

We can see the differentiation and integration of the $\sqrt[4]{x^3}$ on Figure 5. At the end we calculate the numerical value of the definit integral.

In MATLAB we use the functions `int` and `diff` to get the (definit or indefinit) integral or the derivative of the function respectively.

More functions can be used in two form in sympy. As we used `limit` earlier, and as a method of the expression, as we used the `subs`. We can use `integrate`, `diff` and `limit` as function and as a method. Usually the letter is the easier to read, so we use it in such a way. Using as a method we must not give the expression as parameter, so we have one less parameter.

We can see the differentiation and integration of the $\sqrt[4]{x^3}$ on Figure 6. At the end we calculate the numerical value of the definit integral using 26 significant digit.

VI. PLOTTING

Both sympy and MATLAB can plot function given by symbolic values. MATLAB have the `ezplot` function, sympy the `plot` function to plot the figures. Both of these functions have the opportunity to control the parameters of the plot. A plot made by MATLAB and SymPy can be seen in Figures 7 and 8 respectively.

VII. CONCLUSION

Both MATLAB and Python can perform symbolic calculation that needs in a basic calculus course. Both have its advantages and disadvantages. MATLAB has one integrated development environment to do the calculation, while Python

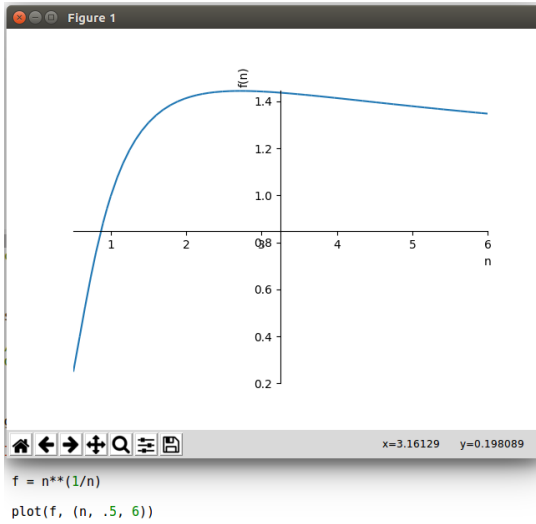


Figure 7. Plot of the $\sqrt[n]{n}$ made by sympy

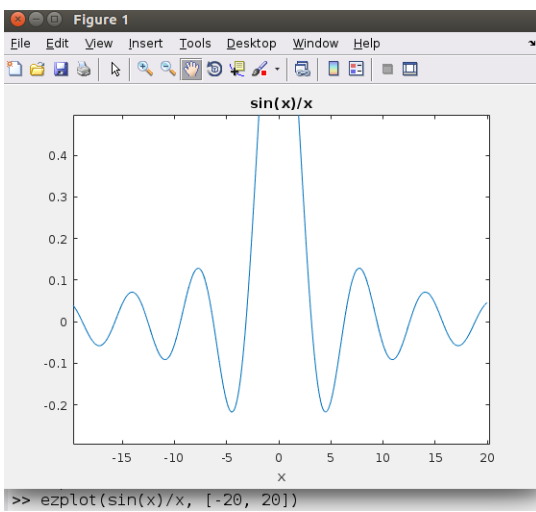


Figure 8. Plot of the $\frac{\sin x}{x}$ made by MATLAB

has several with different advantages. To choose one of them to teach calculus (or other part of mathematics) is can depend on their needs in their later subjects. For example geographer students can use Python for scripting QGIS or ArcGIS program later and achieve geographical databases, so Python can be a good choice to learn as first language.

REFERENCES

- [1] "Official Python site." [Online]. Available: <http://www.python.org>
- [2] M. Summerfield, *Programming in Python 3: A Complete Introduction to the Python Language*. Addison-Wesley Professional, 2010.
- [3] V. István, "Learning and teaching combinatorics with sage," *Teaching Mathematics and Computer Science*, vol. 10, no. 2, pp. 389 – 398, 2012.
- [4] "Official site of Symbolic Math Toolbox." [Online]. Available: <https://ch.mathworks.com/products/symbolic.html>

Collaborative Spatial Keyword Top-k Query

Liang Liu*, Shuai Guo*, Xiaolin Qin*, Qinxue Wang*

* College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

liangliu@nuaa.edu.cn, marvel_agent@nuaa.edu.cn, qinxcs@nuaa.edu.cn, nuaawqx@163.com

Abstract—The proliferation of geo-social network, such as Foursquare and Flickr, enables users to generate location information and its corresponding descriptive keywords. Spatial keyword queries are used to find objects in the geo-social networks. Typical spatial keyword queries meet only a single user's need at a time, which contain a single query location and a single set of query keywords. Collaborative Spatial Keyword Top-k Query (TKCSKQ) asks for top-k objects that are close to multiple query positions and their keywords have high relevancy with multi-group query keywords. To solve the problem that there are repeated and synonymous keywords in multi-group query keywords, a keywords similarity calculation formula based on the weight of query keywords weight is designed. And we propose SKNIR-tree to support near keywords matching, which is an extension of the IR-tree. Based on the SKNIR-tree, we propose a query processing algorithm that prunes search space through maintaining a priority queue and calculating the minimum spatial and textual similarity of each node with the query, to quickly identify the desired objects. Extensive experiments on real dataset validate the efficiency and the scalability of the proposed algorithm.

Keywords: spatio-textual object, spatial keyword query, Top-k query, collaborative query

I. INTRODUCTION

With the wide application of the localization technology, geotagging is incorporated into the text data. For example, photo sharing sites (e.g., Flickr) have many photos which contain the location and text description. As another example, check-ins or reviews in location based social networks (such as Foursquare) contain both text description and locations of points of interest. Spatial keyword query processing technologies [1,2,3,4] are used to identify the desired spatio-textual objects efficiently, which have high relevancy with the query while taking into account both the spatial proximity and the text similarity.

Typical spatial keyword queries meet only a single user's need at a time, which take a single query location and a single set of keywords as input parameters, return the objects that their locations are near the query point and their keywords are highly similar to the query keywords. But in real life, users often draw up a plan collaboratively. The query in these applications is submitted by multiple users. For example, users who are in different companies dine together. Each user wants the restaurant to be near its own location and the restaurant's description is similar to its own need. In this paper, we study how to find suitable top-k objects to meet multiple users' needs. We formulate a new kind of query called collaborative spatial keyword top-k query (TKCSKQ), which aims to retrieve top-k

objects for meeting multiple users' needs. Compared with traditional spatial keyword queries, TKCSKQ faces the following challenges:

(1) There are repeated and synonymous keywords in query keywords submitted by multiple users. Traditional spatial keyword queries take a single set of keywords submitted by a single user as parameter, and there are no duplicate keywords or near keywords in the query keywords. Their keywords similarity calculation method does not consider the weight of the query keywords.

(2) There is mismatch problem because of the synonymous keywords. For example, in a collaborative query, two users propose a query keyword "open-air" and "outdoor" respectively. Clearly, they both want to query outdoor restaurant. For an object that contains "open-air" keyword, it will only match one query keyword while using traditional query processing technology. But in fact, it should match the two query keywords.

(3) How can we process TKCSKQ efficiently? It is another great challenge for TKCSKQ to quickly find top-k objects that are close to multiple query points and their keywords have high relevancy with query keywords.

To solve the above problem, we propose a collaborative spatial keyword top-k query processing technique, and the main contributions are as follows:

(1) To solve the problem that multiple users submit the repeated or synonymous keywords, we design the keywords similarity calculation formula based on the weight of query keywords.

(2) To process TKCSKQ efficiently and solve mismatch problem, we propose an efficient hybrid index structure called Synonymous Keywords Normalization IR-tree (SKNIR-tree), which normalizes all the keywords and uses NKI (Normalized Keyword Identification) to represent keyword, to maximize the users' satisfactions.

(3) Based on the SKNIR-tree, we propose an algorithm TKCSK (Top-K Collaborative Spatial Keyword processing method) that prunes search space through maintaining a priority queue and calculating the minimum spatial and textual similarity of each node with the query, to quickly identify the desired objects.

In order to evaluate the performance of the SKNIR-tree and TKCSK algorithm, we conduct extensive experiments in two aspects of query time and IO. The results demonstrate that the proposed algorithm is efficient and scalable and exhibits superior performance over the brute force method.

The rest of this paper is organized as follows. Section II introduces the related work. We formally define the problem of collaborative spatial keyword top-k query in

Section III. Section IV introduces the SKNIR-tree. Section V introduces our algorithm for processing TKCSKQ. Section VI presents our experimental evaluation. We summarize our work and discuss future work in Section VII.

II. RELATED WORK

Typical spatial keyword queries meet only a single user's need at a time [5,6,7,8]. They mainly construct the hybrid index and propose the corresponding algorithm to search desired objects [9,10,11]. The R*-IF [9] organizes location information with R tree, each leaf node is associated with an inverted file to organize text information. The algorithm finds the nearest neighbor according to the leaf nodes of R tree. And then in each leaf node, the objects are sorted according to the textual relevancy. IR-tree [2] associates an inverted file with each node of the R tree, and uses the priority queue to query objects with the maximum relevancy taking into account both the spatial proximity and keywords relevancy. BR-tree [10] organizes text information through associating a bitmap with each node of R tree. The algorithm prunes search space according to whether the bitmap contains all query keywords. Then the objects are sorted according to the distance. Wu et al. [11] study the authentication of moving top-k spatial keyword queries using the MIR-tree, which modifies the IR-tree by embedding a series of digests in each node of the tree. The above queries are all submitted by a single user. On the contrary, we aim to solve the spatial keyword queries submitted by multiple users and find the desired objects to meet the needs of multiple users.

The existing spatial keyword queries have some collaborative studies, ALI et al. [12] studies the k-BEST-SUBGROUPS-NN query. The query is submitted by multiple users, and asks for results to meet any sub-groups' demand. The algorithm proposes a data centric approach, gradually accesses the objects from centric, and identifies the best subset at each step. The main idea is to develop the best subset of the visited objects by moving the query point radially from the centroid, without enumerating all possible subsets. But this paper only considers the coordination of space, does not consider keywords. Zhang et al. [13] proposes TkCoS query taking into account both the spatial proximity and keywords relevancy, and designs the STR-tree which prunes search space by calculating the upper boundary and the lower boundary for each node set. TkCoS query is submitted by multiple users, and finds the top-k object sets to satisfy the users' needs. The collective spatial keyword queries [14,15,16] are submitted by a single user, and find the top-k object sets. The keywords of each object set contain query keywords, the location of the object set is close to the query location. TKCSKQ is different from the above researches, the query helps multiple users in different locations to identify top-k objects collaboratively while taking into account the problem of the repeated and synonymous keywords.

III. PROBLEM STATEMENT

Spatial textual object. $o = \langle \rho, \varphi \rangle$, where ρ is the object's location, φ is a set of keywords of the object.

Collaborative Spatial Keyword Top-k Query. $Q = \{ \langle q_1 \cdot \rho, q_1 \cdot \varphi \rangle, \dots, \langle q_n \cdot \rho, q_n \cdot \varphi \rangle \}$, where $q_i \cdot \rho$ is the i th user's query location, $q_i \cdot \varphi$ is the i th user's query keywords. TKCSKQ asks for top-k objects that are close to multiple users' locations and their texts are highly similar to the query keywords.

In order to find the best top-k objects from the dataset, we propose a ranking function to measure how well an object satisfies TKCSKQ, as shown in Formula 1. The function takes into account both the spatial proximity and keywords relevancy. In Formula 1, $\alpha \in [0,1]$ is the user preference on spatial proximity and keywords relevancy. The spatial proximity, denoted by $D(Q, o)$, is obtained by the maximum distance between q_i and object (shown in Formula 2). $maxD$ denotes the maximal distance between any two objects in dataset. It is used as a normalization factor. In Formula 3, $TRel(Q, \varphi, o, \varphi)$ is the keywords relevancy. Traditional spatial keyword queries are submitted by a single user, there are no repeated and synonymous keywords in query keywords. So their keywords similarity calculation formula does not consider the weight of a single query keyword. But for TKCSKQ, multiple users may submit the same keywords or synonyms keywords, thus query keywords need to be assigned different weights. We propose a keyword similarity calculation formula based on the weight of query keywords (shown in Formula 3). If t_i represents the keyword that both appear in the query keywords and object keywords, w_i is the weight of t_i , obtained by the number of times t_i 's NKI appears in the query keywords. The smaller the value calculated by Formula 1, the more satisfied the query condition.

Finally, the goal of a TKCSKQ is to find top-k objects with the smallest $S_{sk}(Q, o)$. Our problem can be defined as Definition 1.

$$S_{sk}(Q, o) = \alpha \frac{D(Q, o)}{maxD} + (1 - \alpha)(1 - TRel(Q, \varphi, o, \varphi)) \quad (1)$$

$$D(Q, o) = \max_{1 \leq i \leq n} (\text{dist}(q_i, \rho, o, \rho)) \quad (2)$$

$$TRel(Q, \varphi, o, \varphi) = \frac{w_1 + w_2 + \dots + w_i}{num(Q, \varphi)} \quad (3)$$

Definition 1 (TkCSKQ Retrieval) Given a dataset and a TKCSKQ, find top-k objects $\{o_1, o_2, \dots, o_k\}$, such that there does not exist o' that satisfies $o' \notin \{o_1, o_2, \dots, o_k\}$ and $S_{sk}(Q, o') < S_{sk}(Q, o_i)$, $o_i \in \{o_1, o_2, \dots, o_k\}$.

IV. SKNIR-TREE

Figure 1 is an example of eight spatial textual objects. The left shows the locations of the objects. And the right shows the keywords information, among them, keyword t_2 and t_5 are semantically synonymous.

To answer TKCSKQ efficiently, we introduce an efficient hybrid index structure called SKNIR-tree, which is an extension of IR-tree, as shown in figure 2. It can efficiently standardize the nonstandard keyword into the NKI, to ensure the accuracy and efficiency of the query.

SKNIR-tree normalizes the keywords by maintaining a relational table (shown at the bottom left of Figure 2) and

identifies it with an integer number called NKI. The synonymous keywords will be translated into the same NKI. The application scenarios (e.g., multiple users who are in different corporates dine) TKCSKQ process involve the keywords such as fast food, open-air restaurants and other tabbed keywords, and the number of tabbed keywords is always fixed. Therefore, NKI can be determined in advance, and it is easier to correspond the

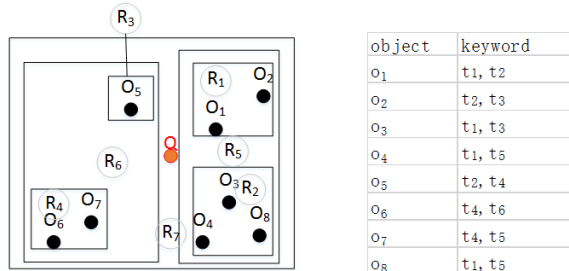


Fig.1 Spatial textual objects

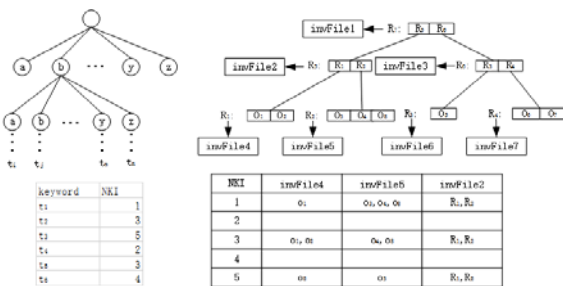


Fig.2 SKNIR-tree

normal keywords to the NKI. Although it is laborious, but only need to be done once. To speed up the query of keywords in the relational table, we use trie (shown at the left above of Figure 2) to organize nonstandard keywords. In the trie, we store an English letter in each node in addition to the root node, each keyword can be accessed through a unique path following its letter order. Each leaf node's form in trie is (key, P), where key is a keyword, and P is a pointer to the keyword in the relational table. During the query process, we start from the root node in trie, then in the root node's sub-nodes, hash technology is used to query the node location of the first letter of the keyword, until finding the last letter. The trie uses hash technology to store and query the location of the nodes, thus the query efficiency is high.

We did a proper transformation based on IR-tree to fit TKCSKQ (shown at the right of Figure 2). In the inverted file, NKI replaces the original object keyword. Each leaf node contains entries of the form (op, o.r, IFp), where op is the pointer to the object o, o.r is the bounding rectangle of o, and IFp is the pointer to the inverted file. The inverted file contains two main components: first, all distinct NKIs appearing in the corresponding objects; second, posting lists for each NKI nki that is a sequence of identifiers of the objects whose NKIs contain nki. Each non leaf node contains entries of the form (nps, r, IFp), where nps is the pointer to the child nodes, r is the minimum bounding rectangle of all rectangles in entries of the child node, and IFp is the pointer to the inverted file.

Figure 2 gives an example of SKNIR-tree for the objects in Figure 1. In the relational table, the nonstandard keywords are standardized into digital tags, and both t₂ and

t₅ are normalized to 3 because they have the same meaning. The R tree is constructed according to the locations of the objects, and the NKI inverted files are constructed for each node.

Here we describe the construction of the SKNIR-tree, as shown in the algorithm 1. The overall NKI is pre created, and then the keywords that appear in the dataset are identified with the corresponding NKI (line 1-5). Then insert the (keyword, NKI) into the relational table (line 7), and insert the nonstandard keywords and their address in the relational table into the trie (line 8). Finally, we call the algorithm Insert in IR-tree [2] to insert the object. It is worth noting that the insertion parameters of the SKNIR-tree are the minimum bounding rectangle of the object and the NKIs (line 10).

Algorithm 1: IndexBuilding(o)

1. for each o
2. for each keyword t in o
3. identifying t with NKI
4. end for
5. end for
6. for each nonredundant t
7. insert (t,NKI) to a relational table
8. insert t to trie
9. end for
10. Insert(MBR,NKIs)

V. PROCESSING TKCSKQ

In this section, two Baseline algorithms are first proposed. Then, based on SKNIR-tree, an efficient algorithm for TKCSKQ processing is proposed.

A. Baseline Algorithms

Baseline 1 Unite Subquery (US). Traditional spatial keyword queries are submitted by a single user, and return the objects that their locations are near the query point and their texts are highly similar to the query keywords. TKCSKQ is submitted by multiple users, including multiple query locations and multi-group query keywords, and return the objects that their locations are near the multiple query points and their texts are highly similar to the multi-group query keywords. Intuitively, a brute force approach is to process each subquery q_i in Q using traditional query processing technology independently, and merge all the results returned by the subqueries. Obviously, this approach will lead to high processing cost. First, the same node will be accessed repeatedly in different subqueries. Second, we need to keep the number of the result of each sub-query sufficiently large, to ensure the merged result contains the top-k.

Baseline 2 First Space Then Text (FSTT). This algorithm uses Formula 3 and relational table to calculate the text relevancy of all objects based on the inverted file. The calculation result is denoted as TRank. Then through extending the method of searching neighbor objects [18], the algorithm incrementally finds neighbors that are closest to multiple users using R-tree, and maintains top-k result through calculating neighbors' spatial textual relevancy based on Formula 1. The algorithm keeps track of the maximum text relevancy in TRank, denoted by MaxT that has not been calculating so far. For a newly calculated object in R-tree, if the combined score

computed from its location and $MaxT$ exceeds k th result object, the Algorithm stops since it is guaranteed that all un-calculated objects will not have a lower score than the current k th result object.

B. TKCSK Algorithm

In this section, we propose the TKCSK algorithm to processing TKCSKQ. The algorithm maintains a priority queue that stores minimum spatial relevancy between Q and SKNIR-tree nodes. The relevancy calculation function is shown in Formula 4. $\min S_{sk}(Q, N)$ is the relevancy between Q and the node N . And the relevancies between Q and the objects in minimum bounding rectangle of node N are all greater than $\min S_{sk}(Q, N)$. We give the formal definition in Theorem 1, and give the proof. The priority queue is arranged from small to large. The algorithm iterates over the elements from the head, and calculates the $\min S_{sk}(Q, N)$ of its child nodes, then inserts them into the queue. If we get an object from the head of the queue, then the object is the one of the top- k . Until we get the all top- k result, the algorithm stops. Other nodes and objects in the priority queue do not need to be accessed and calculated to achieve the purpose of fast pruning search space and improving query efficiency.

$$\min S_{sk}(Q, N) = \alpha \frac{D(Q, N)}{\max D} + (1 - \alpha)(1 - TRel(Q, \varphi, N, \varphi)) \quad (4)$$

Theorem 1 Given a TKCSKQ Q and a node N of SKNIR-tree, and the minimum bounding rectangle of the node N contains the objects os , then $\forall o \in os (\min S_{sk}(Q, N) \leq S_{sk}(Q, o))$.

Proof. $\text{dist}(q_i, \rho, N) \leq \text{dist}(q_i, \rho, o, \rho)$, then $\frac{\min_{1 \leq i \leq n}(\text{dist}(q_i, \rho, N))}{\max D} \leq \frac{\min_{1 \leq i \leq n}(\text{dist}(q_i, \rho, o, \rho))}{\max D}$, $D(Q, N) \leq D(Q, o)$. And because the node N contains all the keywords of the objects os , $1 - TRel(Q, \varphi, N, \varphi) \leq 1 - TRel(Q, \varphi, o, \varphi)$. In summary, $\min S_{sk}(Q, N) \leq S_{sk}(Q, o)$.

The algorithm's pseudocode is shown in Algorithm 2. First of all, we transform the query keywords from multiple users into NKIs, and sort NKIs (line 2-7). Then we calculate the number of time each NKI appears as the weight of the NKI (line 8). The algorithm maintains a priority queue U which is arranged from small to large according to S_{sk} , and firstly the root node of the SKNIR-tree is stored in U (line 9-10). If U is not empty and the number of result is less than k (line 11), the algorithm iteratively checks the first element E in U . If E is an object, it is returned as a top- k result. If E is a leaf node, we compute the $S_{sk}(Q, o)$ of E 's objects and push them into U . If E is a non leaf node, we compute the $\min S_{sk}(Q, N)$ of E 's child nodes and push them into U . (line 12-23).

Algorithm 2: Search(index R, TKCSKQ Q)

```

1. Result  $\leftarrow \emptyset$ 
2. for each  $q_i, \varphi$ 
3.   for each keyword  $t$ 
4.     Q.NKIs.add( $t \rightarrow$  NKI)
5.   end for
6. end for
7. sort(Q.NKIs)
8. Q.NKI.w  $\leftarrow$  count(Q.NKI)
9. U  $\leftarrow$  EmptyPriorityQueue

```

```

10. U.push(R.root, 0)
11. while U is not empty and Result.length < k
12.   E  $\leftarrow$  U.pop()
13.   if E is an object
14.     result.add(E)
15.   else if E is a leaf node
16.     for each object  $o$  in the leaf node
17.       U.push( $o, S_{sk}(Q, o)$ )
18.     end for
19.   else
20.     for each node  $n$  in E
21.       U.push( $n, \min S_{sk}(Q, N)$ )
22.     end for
23.   end if
24. end while

```

VI. EXPERIMENTS

A. Experimental Setting

The experiment is performed on ThinkPad T450, with the following configuration: CPU: Intel (R) Core (TM) i5-5200U CPU @ 2.20GHZ, RAM: 6G, Hard disk: 500G, Operation System: Windows 10. All algorithms of the experiment are implemented in Java, and the integrated development environment is IntelliJ IDEA Community Edition 14.0.2.

We use the yelp_academic_dataset_business dataset [17] provided by the Yelp web site in the experiments. It collects 85,901 restaurants from 11 cities in 4 countries. Each line in the dataset records a restaurant's information which contains 31 items, such as merchant identification, address, latitude and longitude, classification etc. We use latitude and longitude as object location and use classification as object keywords. We also extend the dataset by the method of random sampling based on the original dataset. Because the keywords in the dataset do not have synonymous keywords, we randomly select multi-group 2-4 keywords as synonymous keywords.

The existing technologies about spatial keyword query cannot deal with TKCSKQ. Thus we only compare our TKCSK algorithm with two Baseline algorithms proposed in the 5.1 section. We normalize the object keywords and then put it into memory in FSTT algorithm in advance.

B. Performance Evaluation

We compare TKCSK algorithm with two Baseline algorithms in two aspects of query efficiency and IO cost, and the IO cost is measured by the number of objects accessed. In the following, n is the number of users, k is the number of results, and α is the user's preferences on spatial proximity and keywords relevancy.

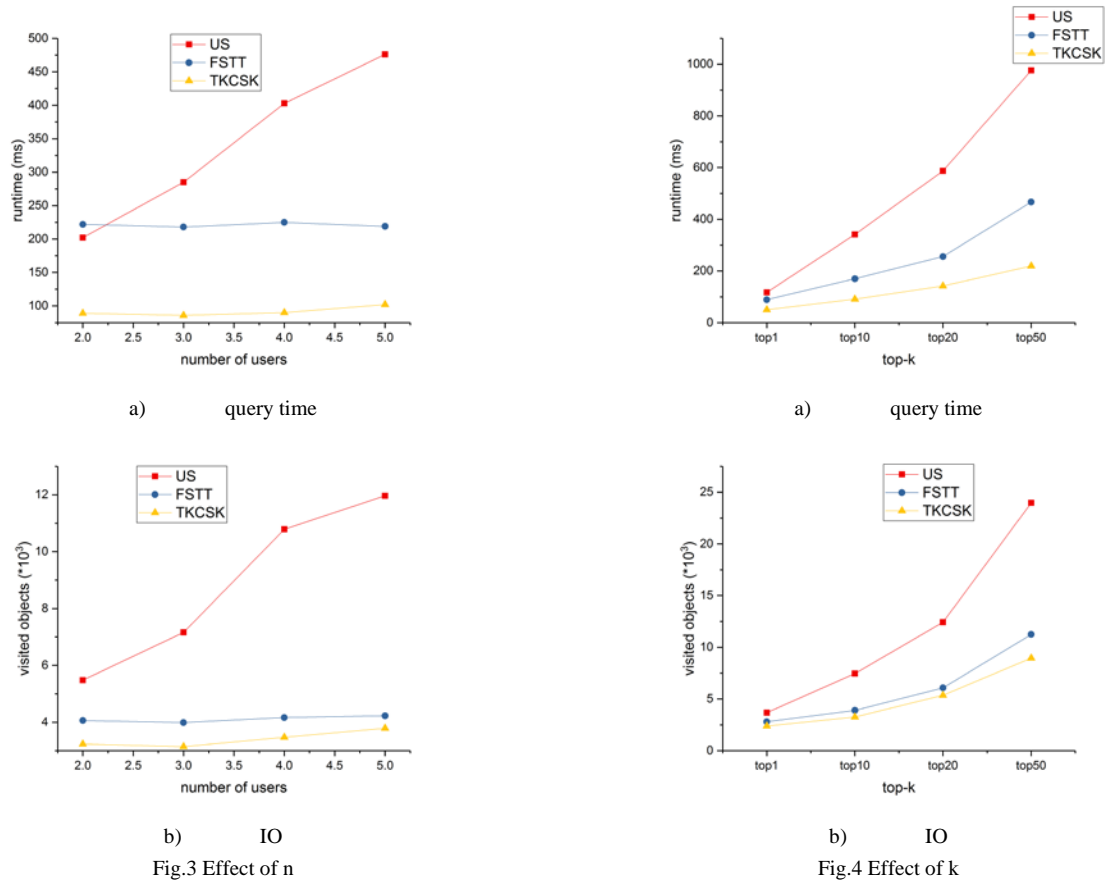


Fig.3 Effect of n

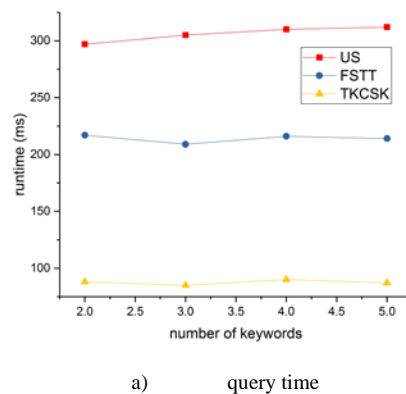
Fig.4 Effect of k

(1) Effect of n

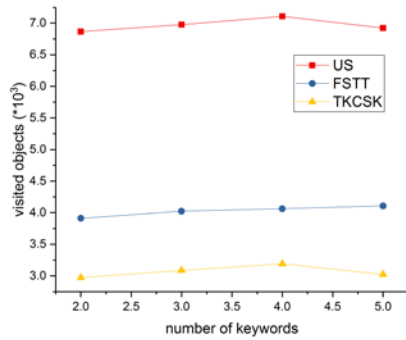
Here, we fix k at 10, the number of query keywords for each user at 3 and α at 0.5. Fig.3 shows the impact of different number of users on query time and IO cost. Because each subquery in the US algorithm will access the index once, and each subquery needs to maintain a result that is much larger than k to ensure the fusion result including top- k , so the query time and IO cost of US algorithm are much larger than that of TKCSK algorithm. In the FSTT algorithm, all the object NKIs are stored in memory, and the algorithm incrementally finds neighbors that are closest to multiple users using R-tree. The algorithm stops when the combined score computed from its location and $MaxT$ exceeds k th result. So the number of object accessed is not much different from that of TKCSK algorithm as shown in Fig.3(b). However, as each iteration step of the algorithm needs to find the $MaxT$ in $TRank$, though it may not be needed every step, but it also causes high cost, so the query time of FSTT algorithm is greater than that of TKCSK. In addition, the number of subqueries increases with the increase of n in US algorithm, so as shown in the figure, both the query time and the IO cost increase. The increase of n does not affect the pruning rate of TKCSK and FSTT algorithms, so as shown in the figure, with the increase of n , the query time and IO cost of TKCSK and FSTT algorithms are almost unchanged.

(2) Effect of k

In this set of experiments, we evaluate the performance of the three algorithms with a varying k while fixing n at 3, the number of query keywords for each user at 3 and α at 0.5. As shown in Fig.4(a) and Fig.4(b), the TKCSK algorithm has shorter query time and smaller IO cost than two Baseline algorithms for all values of k . With k increasing, as the number of results increases, the amount of pruning will decrease accordingly, so as shown in the figure, query time and IO cost increase. And since the US algorithm needs to maintain a larger value than k (experimental setting is 2 times), its growth rate is greater than that of TKCSK algorithm and FSTT algorithm.



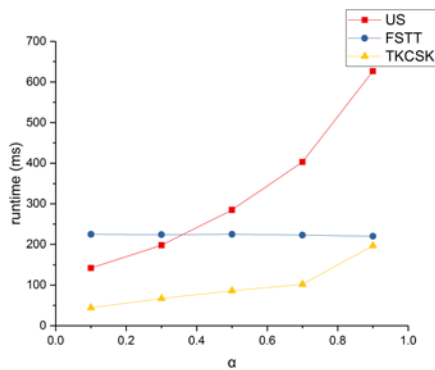
a) query time



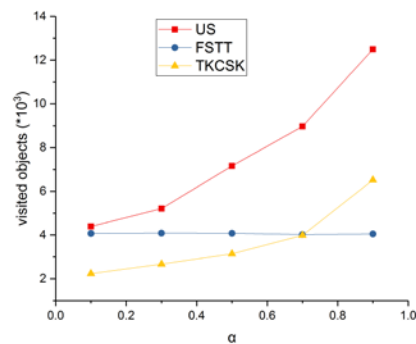
b) IO
Fig.5 Effect of keywords

(3) Effect of the number of query keywords

Fig.5 shows the effect of the number of query keywords for each user on query time and IO cost when we fix k at 10, n at 3, and α at 0.5. Specifically, TKCSK algorithm has shorter query time and smaller IO cost than two Baseline algorithms for all values of the number of keywords. As shown in figure, query time and IO cost remain unchanged with the number of query keywords increasing, because the number of keyword queries does not affect the pruning rate of the all algorithm.



a) query time

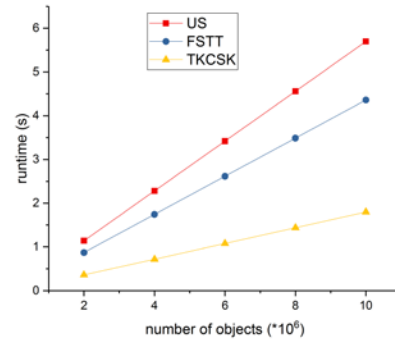


b) IO
Fig.6 Effect of α

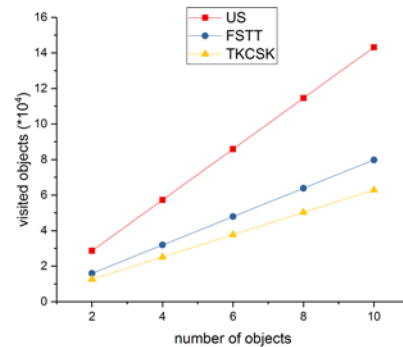
(4) Effect of α

Fig.6 shows the effect of α on query time and IO cost when we fix k at 10, n at 3 and the number of query keywords for each user at 3. Specifically, TKCSK algorithm has shorter query time and smaller IO cost than two Baseline algorithms for all values of the α . Recall that α is used to adjust user's preferences for spatial proximity

and keywords relevancy. The greater α value is, the more user cares about the location of results. The smaller alpha value is, the more user cares about the keywords relevancy of results. As shown in the figure, query time and IO cost of US and TKCSK algorithms increase with the increase of α . This is because the spatial differentiation is small and the pruning rate using spatial proximity is small. For FSTT Algorithm, because it firstly uses R tree to incrementally calculate the object closest to multiple users, the change of α will not affect algorithm's termination condition, therefore query time and IO cost of the FSTT algorithm do not change with α increasing.



a) query time



b) IO

Fig.7 Scalability

(5) Scalability

In order to evaluate the scalability of TKCSK, we generate dataset from two million to ten million based on the original dataset. The location of the generated object is the random neighbors of the location of the object in the original dataset, and the keyword is randomly obtained from the keyword set in the original dataset. Fig.7 shows the tendency of query time and IO cost of algorithms with changing the amount of data when we fix k at 10, n at 3, α at 0.5 and the number of query keywords for each user at 3. As shown in the figure, TKCSK algorithm is scalable and better than two Baseline algorithms.

VII. CONCLUSIONS

In this paper, we study the problem of collaborative spatial keyword top-k query (TKCSKQ), which aims to find top-k objects that are close to multiple query points and theirs texts have high relevancy with query keywords. Because there are repeated and synonymous keywords in query keywords, we design the keywords similarity

calculation formula based on the weight of query keywords. To solve mismatch problem and efficiently process TKCSKQ, we present an efficient query processing algorithm that is based on a hybrid index called SKNIR-tree. The algorithm prunes search space through maintaining a priority queue and calculating the minimum spatial and textual similarity of each node with the query locations and query keywords, to quickly identify the desired objects. Our experimental evaluation shows that the proposed algorithm is efficient and scalable and superior performance compared with two baseline methods.

REFERENCES

- [1] De Felipe I, Hristidis V, Risse N. Keyword Search on Spatial Databases[C]// IEEE, International Conference on Data Engineering. IEEE Computer Society, 2008:656-665.
- [2] Cong G, Jensen C S, Wu D. Efficient retrieval of the Top-k most relevant spatial web objects[J]. Proceedings of the Vldb Endowment, 2009, 2(1):337-348.
- [3] Chen Y Y, Suel T, Markowitz A. Efficient query processing in geographic web search engines[C]// ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, Usa, June. DBLP, 2006:277-288.
- [4] Cao X, Cong G, Jensen C S. Retrieving Top-k prestige-based relevant spatial web objects[J]. Proceedings of the Vldb Endowment, 2010, 3(1):373-384.
- [5] Cao X, Chen L, Cong G, et al. Spatial Keyword Querying[M]// Conceptual Modeling. Springer Berlin Heidelberg, 2012:16-29.
- [6] Chen L, Cong G, Jensen C S, et al. Spatial keyword query processing: an experimental evaluation[J]. Proceedings of the Vldb Endowment, 2013, 6(3):217-228.
- [7] De Felipe I, Hristidis V, Risse N. Keyword Search on Spatial Databases[C]// IEEE, International Conference on Data Engineering. IEEE Computer Society, 2008:656-665.
- [8] Zhang D, Tan K L, Tung A K H. Scalable Top-k spatial keyword search[C]// International Conference on Extending Database Technology. ACM, 2013:359-370.
- [9] Zhou YH, Xie X, Wang C, GongYC, Ma WY. Hybrid index structures for location-based Web search. In: Proc. of the CIKM. New York: ACM Press, 2005. 155-162. [doi: 10.1145/1099554.1099584].
- [10] Zhang DX, Chee YM, Mondal A, Tung AKH, Kitsuregawa M. Keyword search in spatial databases: Towards searching by document. In: Proc. of the ICDE. Washington: IEEE, 2009. 688-699. [doi: 10.1109/icde.2009.77].
- [11] Wu D., Choi B., Xu J., C.S. Jensen. Authentication of moving top-k spatial keyword queries. IEEE Trans. Knowl. Data Eng., 27 (4) (2015), pp. 922-935
- [12] Ali M E, Tanin E, Scheuermann P, et al. Spatial Consensus Queries in a Collaborative Environment[J]. Acn Transactions on Spatial Algorithms & Systems, 2016, 2(1):3.
- [13] Zhang J, Meng X, Zhou X, et al. Co-spatial Searcher: Efficient Tag-Based Collaborative Spatial Search on Geo-social Network[C]// International Conference on Database Systems for Advanced Applications. Springer-Verlag, 2012:560-575.
- [14] Cao X, Cong G, Jensen C S, et al. Collective spatial keyword querying[C]// ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June. DBLP, 2011:373-384.
- [15] Zhang P, Lin H, Yao B, et al. Level-aware Collective Spatial Keyword Queries ☆[J]. Information Sciences, 2016, 378(C):194-214.
- [16] Long C, Wong C W, Wang K, et al. Collective spatial keyword queries: a distance owner-driven approach[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2013:689-700.
- [17] https://www.yelp.ca/dataset_challenge/dataset
- [18] G. R. Hjaltason and H. Samet. Distance browsing in spatial databases. ACM Trans. Database Syst., 24(2):265-318, 1999.

Low-voltage LED lighting system integrated with solar power cell and monitoring of the LEDs

S.Grigoryeva*, D.Titov*, and Gy.Gyorok**

* D.Serikbayev East Kazakhstan State Technical University/ Faculty of Information Technology and Energy, Ust-Kamenogorsk, Kazakhstan

** Obuda University/ Alba Regia Technical Faculty, Hungary
SGrigorieva@ektu.kz, DTitov@ektu.kz, gyorok.gyorgy@arek.uni-obuda.hu

Abstract—An article puts forward the concept of and actualizes a new LED system of interior lighting for administrative buildings. The main advantage of given lighting system, which is also its scientific novelty, is its use of safe low-voltage power supply (24V) for LED lamps, absence of transformers and integration with solar power cells, which allows to significantly bring down the energy consumption of lighting.

I. INTRODUCTION

In accordance with the Paris Agreement on Climate 2016, Kazakhstan committed itself to reducing greenhouse gas emissions by 15% by 2030 compared to 1990 [1]. This means either to save electricity or to switch to alternative energy sources. At the same time there is an industrial development of Kazakhstan, which provides for an increase in electricity consumption. The electricity consumption for lighting is about 12% of total consumption [2]. Therefore, the solution of the diametrical tasks stated above is partially possible with the use of energy-efficient LED lighting systems.

In this paper, we are considering the creation of an energy-efficient low-voltage LED lighting system with serviceability monitoring of LEDs and integration with solar power cells.

Currently, lighting systems are used with high-voltage power - 220V in all countries of the world. The development of new technologies in the field of alternative energy and the production of high-performance, reliable LED crystals of increased power makes it possible to change the approach to organizing the lighting of the building. Fundamentally new element is the coincidence of the power of LED lighting elements and alternative energy sources (solar panels) to 24V, which allows the development of a new architecture of lighting systems with battery life.

The idea of the research is to replace hazardous high voltage 220V in lighting systems for safe lighting 24V, which coincides with the power of energy-efficient LED lamps and the voltage generated by solar cells. This will allow the transition to a fundamentally new architecture of lighting systems. Changing the power supply system will significantly reduce energy consumption due to the use of LED lighting without voltage converters, increase the reliability and durability of the lighting system, and significantly improve the safety of working with the lighting system. A constant current with a voltage of 24V does not pose a danger to human life.

Increasing the safety of the lighting system will completely eliminate injuries from electric current when lighting is used. In turn, this will reduce the cost of health care and the number of deaths of the population.

The classical scheme of switching on the LED in the lighting device is shown in Fig. 1.

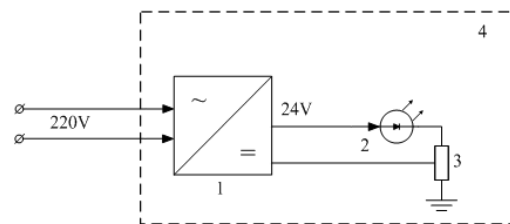


Figure 1. The classical scheme of the LED lighting device

Power supply 1 (converter from AC to DC) is one of the elements of the lighting device. Loss of energy when lighting a room by LED lamp connected to a network with alternating voltage of 220V is wasted with voltage conversion.

Wide introduction of alternative energy sources, such as solar cells or solar cells for the use of typical lighting devices leads to the need to convert the DC current received from them to AC 220V. Usually, when using alternative energy sources, lighting devices are used that are schematically represented as a circuit 4 (Fig. 1). To provide AC voltage 220V converter is installed 24-220V. In this case, the general scheme for implementing the connection of LED lighting systems is shown in Fig. 2.

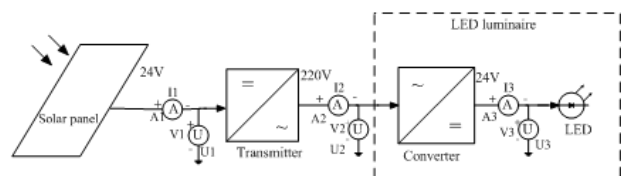


Figure 2. General scheme for system research

Despite the low efficiency of solar panels, their introduction is promising due to the use of renewable energy. However, in the scheme under consideration there are two transducers - positions 2 and 3, which introduce losses and leakages into the lighting system. The first transmitter 2 converts the DC current generated by the solar battery 24V into the working voltage of the AC 220 V network, and the second transmitter 3 converts the AC into a constant current from 220V to a voltage of 24V.

II. EXPERIMENTAL RESEARCH AND RESULTS

The use of low-voltage lighting system will save energy lost by double conversion of current [3,4]. Design has been collected consisting of a solar panel, a transmitter, a converter, an LED instrument, a voltmeter, an ammeter for studying the energy losses was collected (Fig. 2).

Calculate the efficiency of this scheme. The values of the measured currents and voltages are given in Table 1.

TABLE I
EXPERIMENTAL VALUES OF CURRENTS AND VOLTAGES

I_1, A	U_1, B	P_1, W	I_2, A	U_2, B	P_2, W	I_3, A	U_3, B	P_3, W
3,5	24	85	0,35	220	77	2,7	24	64

The current and voltage are measured before and after the converters. These data calculate the power before and after the converters and the efficiency of both converters and the total efficiency of the system. The results of calculations based on the experimental values obtained are shown in Table 2.

The obtained results showed ineffective application of double voltage conversion with loss of energy by 30%.

TABLE II
ESTIMATED VALUES OF ENERGY EFFICIENCY

P_1, W	P_2, W	P_3, W	$\eta_1 \%$	$\eta_2 \%$	$\Sigma \eta \%$
85	77	64	90	80	72

In the absence of converters, the question arises of stabilizing the current. If we use 20 "Amstrong" luminaries on the floor of an office building, we need to provide a current of the order of 20A at a voltage of 24V. In the event of failure of one lamp with parallel switching, there will be redistribution of currents on the LEDs. The current on the remaining lamps will increase, which will lead to gradual degradation of the LEDs.

We suggest using a source with voltage stabilization for low-voltage power supply of LED lighting. In this case, if the LED lamp fails, the current will also be redistributed along the lamps, but the current in the lamps will decrease. This will reduce the illumination on the floor.

For research the operation of LED lighting systems, we have developed two schemes - classical, using drivers with current stabilization (Fig. 3) and a circuit with voltage stabilization, including the control of LED failure (Fig. 4).

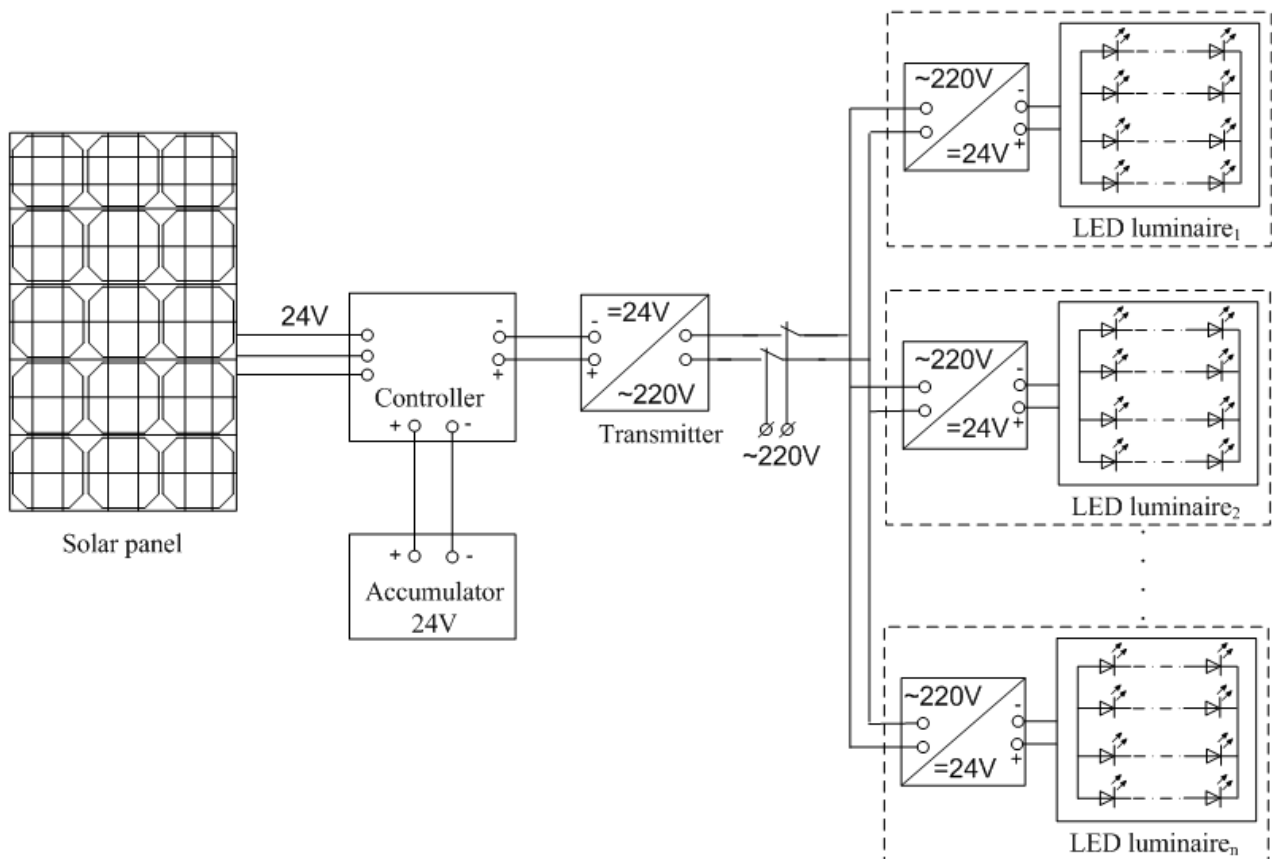


Figure 3. Classical scheme of LED lighting with current stabilization

The classic connection of LED lighting fixtures uses voltage converters. This scheme was implemented on the floor of the educational building of the East Kazakhstan State Technical University [5].

For power supply, two flexible solar panels with a capacity of 450W were used (the total maximum power of

two panels 900W). The corridor was illuminated by 20 luminaries 32W (the total power consumption 640W). Four batteries with a capacity of 650A·h with a voltage of 12V were used in the lighting system. Two batteries were connected in series with total voltage of 24V.

The controller for the operating mode of the batteries and the matching of the external circuit to the solar panel was standard (PWM controller 40A). The 24/220 voltage converter was taken from a standard uninterruptible power supply. The scheme provides for switching to the internal system of power supply 220V when the solar battery fails.

Fig. 4 shows applying of the LED lighting system in accordance with the described scheme.



Figure 4. LED lighting of the floor educational building

The developed control and monitoring system [6] provided data on the energy efficiency of LED lighting using drivers with current stabilization and a solar cell. This lighting system allows you to save electricity almost 10-12 times compared to lighting based on energy-saving fluorescent lamps installed on other floors of the university building.

Integration of the solar power source into the LED lighting system allows to exclude the process of voltage conversion and to provide an operating voltage of 24V.

We are invited to monitor the operation of each lamp in real time to maintain the operating mode of the low-voltage lighting system. The ammeter is used as a sensor to detect a failure. With this control, the circuit of a low-voltage LED lighting system integrated with solar cells is shown in Fig. 5.

Consider the operation of the proposed circuit using the example of a single lamp. The LED matrix consists of 4 LED strips, each of which includes 8 series-connected LEDs. Power to the LED matrix can be fed from a power supply or solar panel. Since the LED strip located in the matrix is designed for 24V, we have the opportunity to use the solar panel as a power source, which saves energy.

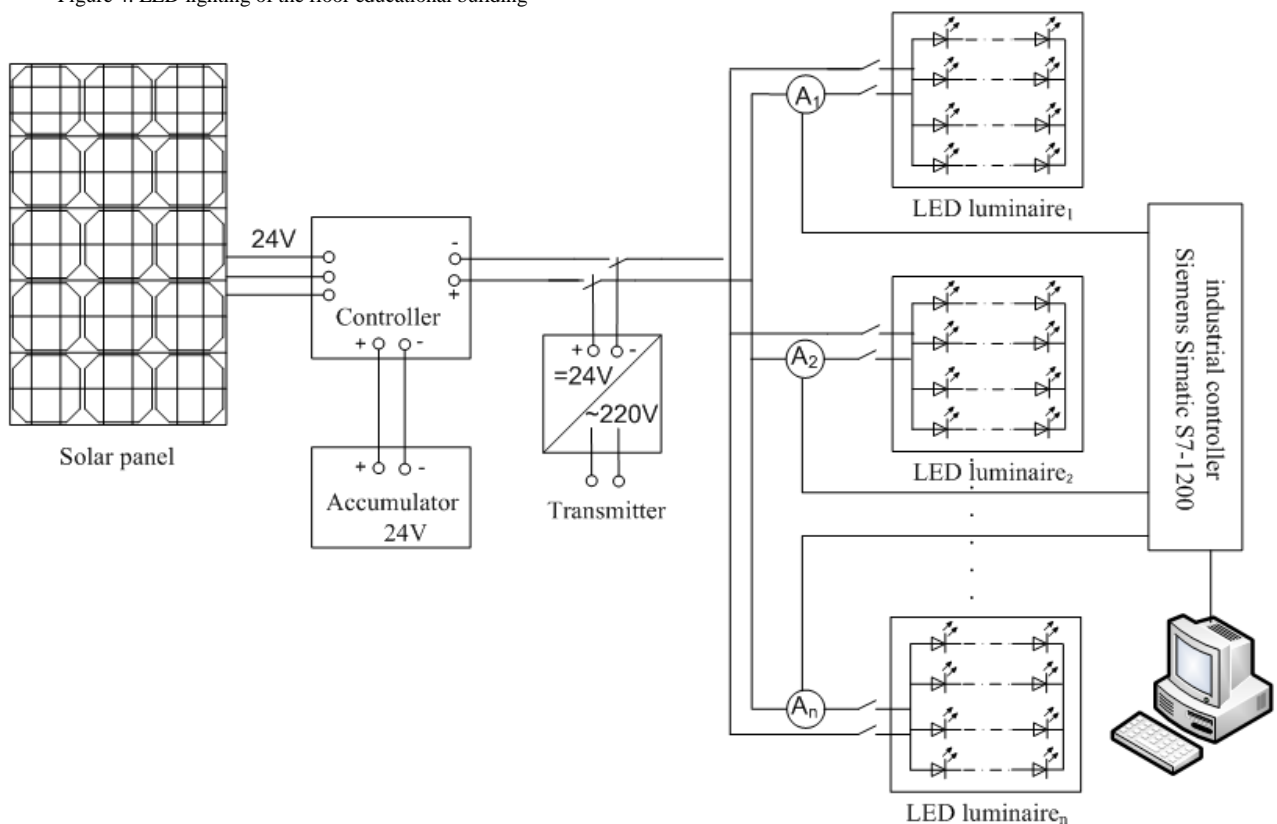


Figure 5. Classical scheme of LED lighting with current stabilization

A sensor connected to the power supply unit measures the level of current that the LED matrix consumes. Next, the measured current level value is sent to the Siemens Simatic S7-1200 controller (Fig. 6).

After that, the controller compares the received data with the set value (a certain constant set for each system individually).

Visualization of the modes of operation of the LED lighting system is shown in Fig. 7.

If the received data is less than set value, the information on the failure of the number of LED ribbons in the corresponding matrix is displayed on the personal computer and on the touch screen of the controller Siemens (Fig. 7,b).



Figure 6. The appearance of the industrial controller unit

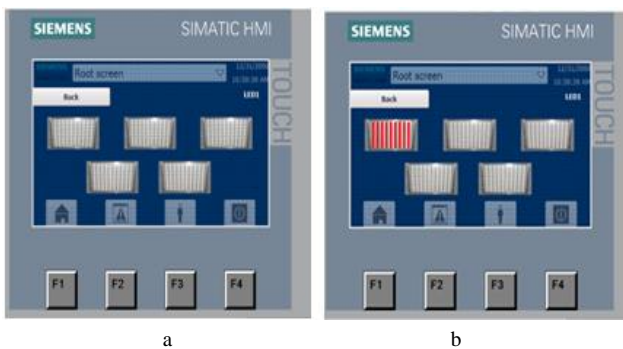


Figure 7. Operational information on work and fault of the LED lighting system: a - all LED lamps work normally; b - the lamp of module 1 failed

If the value of the received value is greater than the set value, it means that a short circuit has occurred in the system. This can damage the power supply if it does not have short-circuit protection and all LED matrixes go out. In addition, there is a high probability of a fire. Therefore, we propose to install a solid state relay on the LED matrix and connect it to the controller. The controller will control the LED matrix via a relay. This will disconnect the LED matrix in the event of a short circuit, which ensures the safety of the system.

CONCLUSION

The use of a low-voltage lighting system allows the integration of solar cells into the lighting system. The possibility of using solar energy makes the system energy efficient (Fig. 8).

Energy efficiency data were obtained experimentally as a result of research of two LED lighting systems (Fig.3,5).

The histogram shows the level of energy consumption in 4 cases:

- 1 - with the converter from the network 220V;
- 2 - from the 220V network without a converter;
- 3 - together with a solar panel with a converter;
- 4 - only the solar panel 24V.

It is seen that about 30% of the energy goes to the work of the converters.

In this case, using solar energy consumption goes only emergency lighting and maintenance work, and is on average about 50 kW·h.

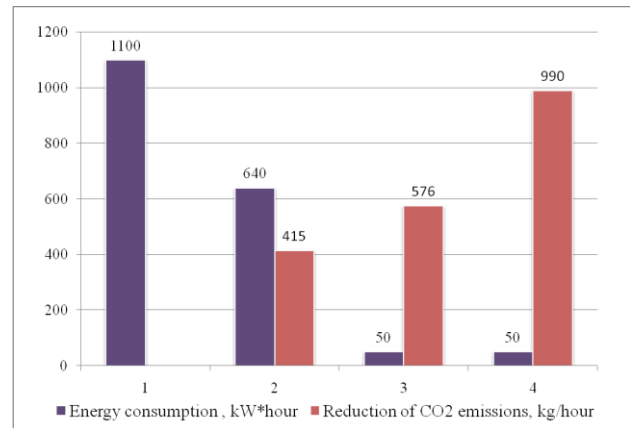


Figure 8. Results of saving electricity of low-voltage lighting system

The use of alternative energy sources in the proposed lighting system will not only save energy, but also reduce consumption of non-renewable energy resources. And as a result, improve the environmental situation associated with increased volumes of hydrocarbon fuel emissions.

If we calculate the energy generated by the solar panels and convert it into an equivalent amount of CO₂ gas ejected into the atmosphere by the power plant [7], we get a reduction in emissions by 990 g/h. When lighting the corridor during the year, an average 4 hours will reduce emissions of the order of one ton of CO₂.

In conclusion, we note that the proposed new system of low-voltage power LED lighting integrated with solar sources of energy is safe, cheaper (due to the lack of converters), more reliable and energy efficient. The results obtained make it possible to draw conclusions about the prospects and possibilities for the practical implementation of the basic principles of low-voltage lighting.

REFERENCES

- [1] Kazinform International news agency. Kazakhstan signs Paris Climate Agreement: http://www.inform.kz/ru/kazakhstan-podpisal-parizhskoe-soglashenie-po-klimatu_a2932653
- [2] Babko A., Inyutin S. Energy and light audit in buildings, structures and street lighting: textbook // Astana: Publishing House ..., 2014. – p.174.
- [3] Baklanov A., Zhaparova A., Titov D. Improving the Efficiency of Led Lighting by Switching to Low-voltage Technology // International Conference on Industrial Engineering (ICIE-2015), 22-23 October 2015, Chelyabinsk and Novocherkassk, Russia.
- [4] Baklanov A., Zhaparova A., Titov D., Gyorok Gy. Study of the Effectiveness of Switching-on LED Illumination Devices and the Use of Low Voltage System in Lighting // Acta Polytechnica Hungarica. – 2015. – Volume 12. – P.71-80.
- [5] Grigoryeva S., Baklanov. A., Titov D., Sayun V., Grigoryev Ye. Analysis energy efficiency of automated control system of LED lighting. // IEEE-Euroasian Conference on Future Energy and International Siberian Conference on Control and Communications (SIBCON–2017), June 29-30,2017, Astana, Kazakhstan.
- [6] Grigoryeva S.V., Dmitrieva T.S., Titov D.N. Evaluation of energy efficiency of LED lighting systems by the example of an educational building // Vestnik KazNRTU. – 2017. – №5(123). – P. 273-277.
- [7] Calculating CO₂ emissions from a coal-fired power plant // The Institute for Energy and Environmental Research – 2016. – Volume14 (3). – P.11-15.

The reconstruction of the medieval Hungarian standard of length based on the geodetic survey of contemporary buildings

György Busics* and Sándor Tóth**

* Óbuda University Alba Regia Technical Faculty, Székesfehérvár, Hungary

** Interplaninfo Ltd., Szeged, Hungary

e-mail busics.gyorgy@amk.uni-obuda.hu; toth.sandor.geo@gmail.com

Abstract — As it is known from documents of the Middle Ages there was an independent unit of length in Hungary between the 11th and 16th centuries – the etalon of that was kept in Székesfehérvár. The 1/16 part of the so-called king's length (royal fathom) was published in statute books. The etalon of this unit does not exist, only a cord was found the length of which is 3.126 metres. In this article we want to demonstrate that this etalon was used for building churches at that time because the measures of the buildings correspond to an ancient unit. Especially the round churches (rotundas) are interesting from this point of view. We precisely measured some Hungarian medieval round churches, but only three of them will be presented in this paper (Kallósd, Bagod and Ják). We used total stations without prisms, angle and distance measurement for detailed polar survey. The measures of these buildings can also be obtained by precise methods, for example the radius of the circle with adjustment. We redrew the floor plan of these buildings. The measures were first given in metres but later in the ancient unit of length, in Hungarian royal foot. This floor plan was used to recalculate the size of the royal foot in metres. We got 31.9 centimetres for royal foot from three buildings. It means that the Hungarian royal fathom (10 feet) equals 3.19 metres instead of the 'official' value of 3.126 metres. Our assumption and the reconstruction method to obtain conversion factors were thus confirmed.

I. INTRODUCTION

As a result of standardisation, distance data, i.e. lengths are given uniformly in a metre-based system all over the world. Today, the definition of the metre, a unit of length as such, is traced back to the wavelength of light. This definition was recognised by the international association after a Hungarian physicist, Zoltán Bay. In the beginning, the etalon was made in the shape of a metal bar. This platinum-iridium bar with a special cross-section is currently kept in Sevres, France.

Before the metre system was introduced, the Vienna fathom had been the official unit of length in Hungary, like in all the countries within the Habsburg Empire. The etalon of the Vienna fathom can be seen both in Vienna and in Bratislava even today.

Certificates confirm that there used to be an independent Hungarian length measurement system in medieval Hungary. Its etalon hasn't survived. Moreover,

the memory of its existence has since then disappeared from the common knowledge, too.

This article presents the geodetic measurements which led to our attempts to restore the medieval Hungarian standard unit of length. The fundamental idea behind our work is that large buildings were designed and constructed on the basis of architectural plans even in the Middle Ages, which must have been carried out with the aid of the then units of length. We can also reasonably assume that the size of buildings were mostly given in round multiples of the measurement. If we manage to determine the measurements of a wisely chosen building that has been preserved in its original form, we may get the original unit of length in a metric system. Round churches which were built in the 10th century in numerous settlements in the Carpathian Basin are particularly suitable for the subject of such geometric, floor plan analyses. In our article we try to reconstruct the length etalon by accurately determining several measurements of three round churches located in Hungary.

II. THE MEDIEVAL HUNGARIAN SYSTEM OF LENGTH

A. *The names and conversion factors of the medieval units of length based on the archives*

We know little about the history of the Hungarian units of length used in the Middle Ages. Their emergence must have been influenced by the Greek, Roman and eastern cultures. It is likely that these units of length emerged from the actual sizes of human nave parts, which their Hungarian and English names also suggest.

The smallest natural unit of length is the **finger** (Hungarian: 'ujj'), which corresponds to the width of an index finger or the overall width of 4 barley seeds placed widthways next to one another. It was referred to as 'daktylos' by the Greeks and 'digitus' by the Romans. The Greek finger measures 19.3 mm in today's metric system.

The **inch** (Hungarian: 'hüvelyk') corresponds to the width of a man's thumb. It was used all over Europe and it still exists in the systems of measurement of several countries. An inch is equal to 12 lines.

The **palm** (Hungarian: 'tenyér', Latin: 'palmus') is a unit of distance that corresponds to the width of 4 fingers.

The **foot** (Hungarian: 'láb', Latin: 'pes') is a unit that has Greco-Roman origins. It doesn't correspond to the average length of a human foot but 16 fingers or 12

inches. Its length varies from country to country between 27-35 cm, according to today’s metric units.

The **span** (Hungarian: 'arasz', Latin: 'spitama') has two types: the great span is the distance between a grown-up man’s extended little finger and thumb, whereas the little span is the distance between the index finger and the thumb. The Hungarian span (great span) makes 10 fingers.

The **cubit** (Hungarian: 'rőf', Latin: 'sing') probably derives from the length of the forearm. It corresponds to 2 feet, 8 palms or 32 fingers in the Hungarian system.

The **step** is likely to originate from the average length of a step. It makes 3 feet in the Hungarian system.

The **fathom** (Hungarian: 'öl', Latin: 'orgia, cubitus') comes from the distance of a grown-up man’s extended arms. The English and German or Austrian fathom measures 6 feet. However, the Hungarian royal fathom makes 10 feet, i.e. it is much longer than the aforementioned ones. The Hungarian fathom is equal to 5 cubits or 16 inches.

We know the conversion factors above thanks to the research by István Bogdán [1] and they are displayed in Table 1.

its actual size on the side of the page. Although this statute book was reprinted in Leipzig in 1488, then in 1490, and a copy has survived, the length of the royal span cannot be measured. Sadly, the top edge of the relevant page was eaten by mice, the bottom edge was cut off by the binder’s knife. The remaining copies of the second edition suffered a similar loss as the end of the line representing the span was cut off while they were being bound.

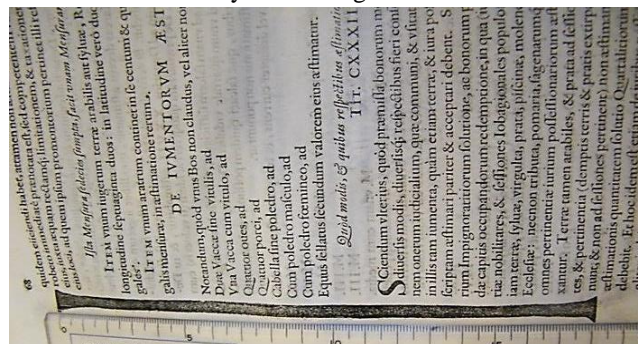


Figure 1. The royal span in the Tripartitum published in Vienna in the year of 1628

TABLE I.
THE MEDIEVAL HUNGARIAN UNITS OF LENGTH AND THEIR EXCHANGE FACTORS

	fathom	step	cubit	span	foot	palm	inch	finger
1 fathom	1	10/3	5	16	10	40	120	160
1 step		1	1,5	24/5	3	12	36	48
1 cubit			1	16/5	2	8	24	32
1 span				1	10/16	40/16	7,5	10
1 foot					1	4	12	16
1 palm						1	3	4
1 inch							1	4/3
1 finger								1

It is also István Bogdán who collected the excerpts from medieval certificates and archival documents, which mention the use and the regulation of units of length [1]. These texts prove that back then there used to be a system of length in Hungary and it was applied, indeed. Some Latin examples include the certificate written by the Chapter of Pécsvárad in 1270 (*'amplexus... cum mensura regia'*), the certificate written by the Chapter of Pécs in 1278 (*'ulna seu mensura... regis et regni'*) or that of the Chapter of Székesfehérvár written in 1368 (*'cubitus seu mensura regalis...'*). Occasionally, you can come across the Hungarian terms, such as in the certificate written by Palatine Miklós Garai in 1379 (*'spatium longitudinalis regalis mensura vulgariter Kyralymertek voce reperissent...'*).

B. The metric length of the royal span

The Hungarian Royal units of length and area were first mentioned in King Matthias’ statute book, and there were drawings as well because the royal span was displayed in

The later statute books, which are known as the Tripartitum by Werbőczy (Hungarian: Hármaskönyv) and of which 50 editions were made, the length of the royal span can be measured using a millimetre ruler. The various editions have been studied by many but their results differ significantly. It is little wonder, though. The paper could have become dry and the printing mould cannot have been perfect either. If we wanted to determine the length of the royal fathom from the size of the span above, we would get a value between 2.88 m and 3.07 m. We must come to the conclusion that this way the fathom cannot be determined precisely enough – it is for informational purposes only.

The royal span was released in the statute book due to the regulation (or standardisation, as we would now call it) of the area measurement. According to the legal text, the royal fathom is equal to the royal span times 16. The unit of area measurement is the royal jugerum-sized land (Hungarian: királyi hold), which is equal to the area of 12×72 royal fathom.

C. The metric length of the royal fathom

There are certificate data in Latin about measuring ropes (measuring cords) used in the Middle Ages to mark and survey areas, which were regarded as an official measure because they had to be transported in a sealed bag. We do not know how long a measuring rope was. We can assume that it was 12 fathoms (maybe 24) long. The width of a royal jugerum-sized land is 12 fathoms, so the rope had to be laid down only once to measure the width and six times to measure the length. The 12-fathom measuring rope is approximately 38 m long – this is similar to the modern-day measuring tapes, which are 20, 30 and 50 m long.

The fact that the royal fathom used to have an etalon (standard measure), which was kept in Székesfehérvár, is known from a certificate that has survived in the archives at the Pannonhalma Archabbey. The Royal Basilica of Székesfehérvár (which has only few remains left to be seen) was the Hungarian kings' coronation and burial site from the time of King Stephen to the Ottoman rule (like Westminster in England, Saint Denis in France, Aachen in Germany and St Vid in the Czech Republic). The crown jewels, the treasury, the archives and the length etalon were all guarded in the provostry that belonged to the basilica.

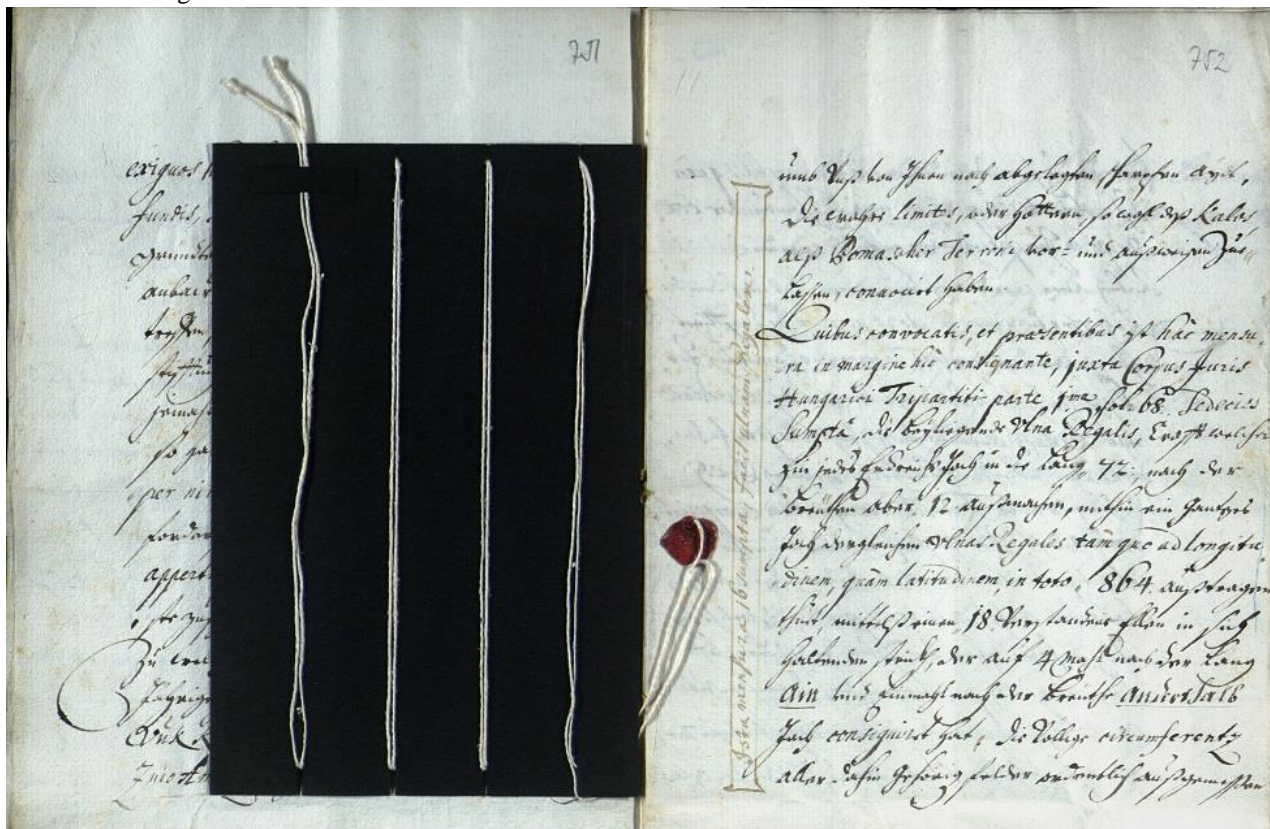


Figure 2. The wound cord and the drawn span in a report from the year of 1702. The place where it is kept: The Hungarian National Archives. Mark: MNL OL E 117 – Fasc. 14. – No. 1.

The above-mentioned certificate is about a land debate between Bakonybél Archabbey and squire of Bakonyszücs. If one of the partners does not regard the area measured by a measuring cord as legal, then they should go to Székesfehérvár (Alba Regalis) and fetch the standard measure of the royal fathom as a heredity of Saint Stephen. The Latin text: *'si mensuram ambiguitatis propulsivam et certam idem dominus abbas habere voluerit, ex tunc hominem suum cum homine eiusdem magistris in Albam Regalem pro aportanda mensura per Sanctum Stephanum regem derelictam et constitutam deberet destinare, alio autem modo nullam iteratam mensurationem acceptaret'*.

Unfortunately, no tangible memory of the etalon has survived, we know of one single copy to be precise, which turned up in the Hungarian central archives. This copy is a royal-fathom long measuring rope, which was attached to a report from 1702 year. Furthermore, a unit of length corresponding to one span was drawn in the report. The

report was written about the survey of the lands that belonged to the two villages. The length of the measuring rope (the distance between the knots tied at the two endpoints of the rope) was measured in the Hungarian Metrology Office and it was said to be 3.126 m. This value is recognised as the metric length of the medieval royal fathom. If the metric value of smaller units is derived from this, we get the following: 1 foot = 31.26 cm; 1 span = 19.54 cm. The latter is in harmony with the distance drawn in the report, which was 19.6 cm. (If the length of the fathom is 3.20 m, the sixteenth of this is 20 cm. The official fathom-span ratio and the determined one are the same. This means that the extent to which the string shrank is equal to the extent to which the drawing on the paper shrank too.)

III. THE POSSIBILITY OF USING BUILDING MEASUREMENTS TO RESTORE THE UNIT OF LENGTH

A. Contemporary buildings as the possible guardians of the unit of length

Contemporary buildings (churches, castles, mansions) are objects usually made of symmetrical geometric shapes that have regular floor plans. We can rightly assume that the marking and the construction of these buildings required the use of plans or else they can't have been built in such quality.

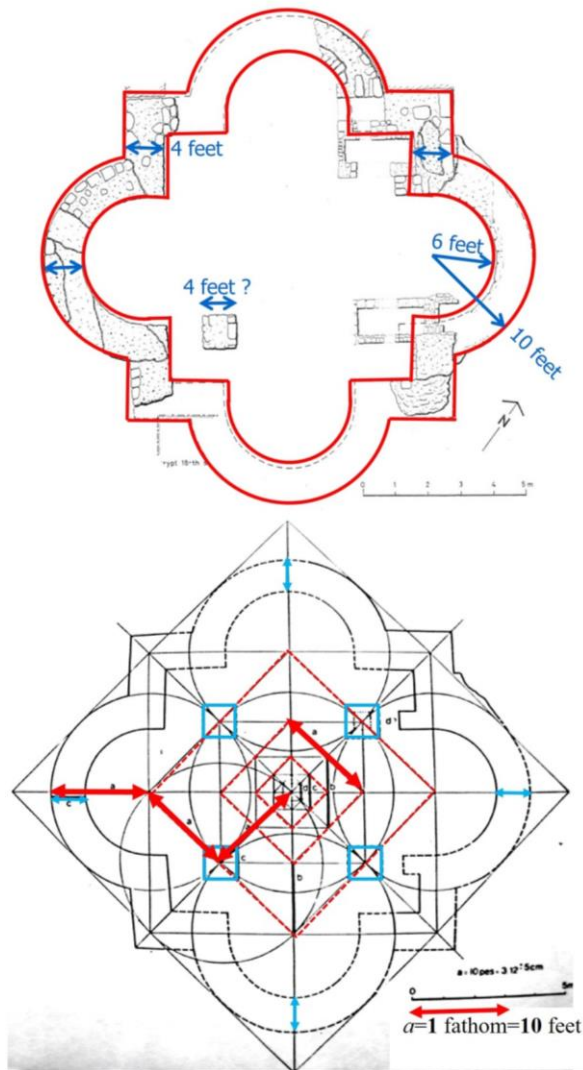


Figure 3. Archaeological floor plan and editing of Prince Géza's church in Székesfehérvár [2]

Architectural design presumes the use of some scale, i.e. the correspondence between a drawing and a real (aerial) unit. It also presumes a system of length. Nowadays, the standard scale of architectural design when using a metric scale is 1:50 (1 mm on the drawing corresponds to 50 mm in reality) or 1:100 (1 mm on the paper represents 10 cm). At the time of the Vienna fathom the scale was 1:72 (1 Vienna inch corresponded to 72 inches in reality, i.e. 1 Vienna fathom or 6 feet). We do not know what the design scale was in medieval Hungary because there is no information about plans that have been preserved intact. It might have been 1:80 for instance, when 1 finger would correspond to 5 feet (0.5 fathom),

but it might have been 1:32, when 1 finger would represent 2 feet.

We can also assume that during the design and the construction the key measures of buildings were provided in round multiples of the unit of length. This simplifies work and is advantageous for practical reasons.

To prove our previous assumption, we studied the measures of 27 medieval churches. They were measured on the basis of archaeological and architectural plans that had been made in the course of heritage preservation. We found that by converting the metric values according to *Table 1* we mostly got round numbers, which made us strongly believe that the former unit of length had been employed.

By way of illustration, we want to share the floor plan of one building and its measures expressed as royal feet with you. This very building was once situated at the highest point of Székesfehérvár and it was the oldest church in the city. It had been built by the first Hungarian king's father, Prince Géza, probably as a chapel. It used to be a church with four vaults, its remaining base walls were excavated by archaeologist Alán Kralovánszky only in 1971 [2]. He reconstructed the design and building process of the regular church and concluded that the outer radius of the vaults corresponded to exactly 1 royal fathom.

For further observations let us reverse the way of thinking detailed above. If the assumption that objects were constructed using the round (or half, maybe one-fourth) multiples of the former unit of length based on plans turns out to be right, then the metric value of the contemporary unit of length could be calculated from the measures of the building based on an accurate survey (carried out in a metric system). It does matter, however, what sort of a building we choose and what method we employ to conduct the survey.

B. The significance of round churches in terms of size determination

Round churches are worth the attention for several reasons. In numerous countries all over Europe, especially in Central-Europe, the oldest churches of the 9th and 10th centuries were built to be circular and quite of few are still standing.

The circle is the simplest geometric shape. It was not only easy to draw with the aid of bows (rondure) on a piece of paper but it was possible to setting out during field work – you needed a string and two poles.

Round churches are advantageous in terms of size determination because in the simplest case there are at least two circles available to be studied – the circle of the outer wall and that of the inner wall. It is also very likely that the thickness of the wall is a multiple of the measure. If the thickness of the foundations and that of the wall are not identical, presumably the difference can be expressed as the former unit of length, too.

Nevertheless, the majority of round churches do not have a base that consists of two circles because the sanctuary and the nave are also rounded. Sanctuaries facing east can also be semicircular, horseshoe-shaped or they have corridor links.

We talk about round churches closing in a semi-circular sanctuary when the centre of the church's arc lies on the

inner arc of the nave (Fig. 4). On the floor plan of such regular churches the radius (R) of the sanctuary is generally half of the radius of the nave, the wall thickness (F) is identical. It is also worthwhile to measure the inner (B) and outer length (K) of the church. Since we get a distance that is longer than the unknown radius ($B=2,5R$; $K=2,5R+2F$), we can determine the radius more accurately. If we expect the R and F values to be round multiples of units smaller than the royal fathom (foot or span), the length of this smaller unit expressed as metres can be determined on the basis of B and K .

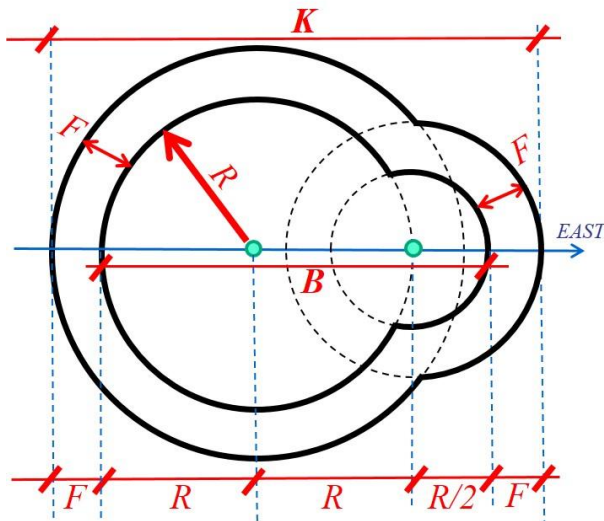


Figure 4. Semi-circular sanctuary closures

We talk about horseshoe-shaped sanctuary closures when the centre of the church's arc lies on the wall or the outer arc of the nave (Fig. 5). The wall thickness of the sanctuary is often half of what the nave possesses.

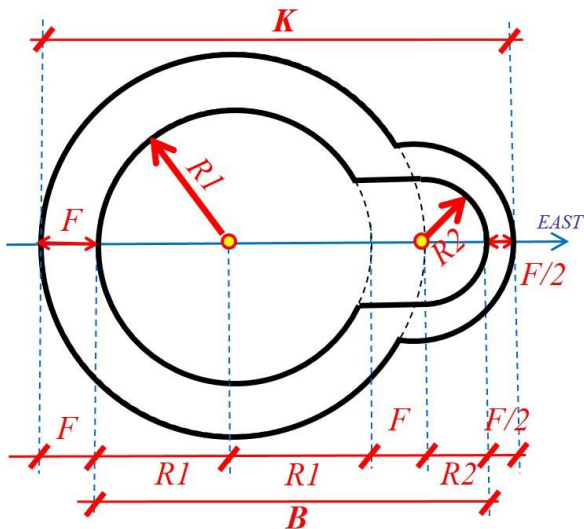


Figure 5. Horseshoe-shaped sanctuary closures

A sanctuary is regarded as elongated when a quadrilateral corridor link is found between the sanctuary and the nave.

A special group within round churches includes those with four vaults. We can determine not only several

circles but also additional things because the location of the vault centres shows regularity.

C. The general technology of building survey – the role of surveying

Since we conducted the survey of numerous round churches as well as the determination of their measures recently, we can propose a technology how etalon reconstruction should be done by generalising our experience.



Figure 6. Good examples to right identify ground wall points at Ják (rough walling) and Kallósd (brick walling) church

- 1.) Choosing the right building to be analysed. This means that it is recommended to choose such a contemporary building that has survived in its original form, the foundation walls are easily identifiable, and the building itself is symmetrical and geometrically regular. It is quite difficult to find a building that has been preserved in its original form because it could have been necessary to rebuild the building to a certain extent over the centuries. In this case, the individual building sections of the different construction periods must be separated as much as possible. The criterion of being identifiable is fulfilled if the foundations or the walls have been built of bricks or ashlar, for instance (Fig. 6). The identification of walls (surfaces, corners) made of rubbles is not clear, therefore such buildings are less ideal for our purposes. Asymmetrical and irregular buildings are unsuitable, too – the circles are ellipses, the columns are not the same, the column intervals are different or there is no other sign of regularity.
- 2.) Identifying the main lines and points of the building to be measured. The key question of all surveys is what we intend to measure. In our case it is enough to survey the elements that comprise the base geometry of the building. To do this, however, we need to study the construction history and the structure of the building or else we cannot choose the right points to be measured.
- 3.) Choosing the right measuring technology. There are different kinds of measuring equipment and technology which may meet our needs: measuring tapes, total stations, laser scanners, UAVs and photogrammetry, etc. For our surveying work we chose the technology of total stations – we presume its use from now on. It is a great advantage that inside and outside the building an accurate geodetic control point network can be created with the aid of direction and distance measurement, the scale of which (its metric system) is provided by the

frequency of the calibrated distance measuring instrument. It is also very practical that we can decide which points we wish to measure, i.e. we need to measure the points only which we find essential.

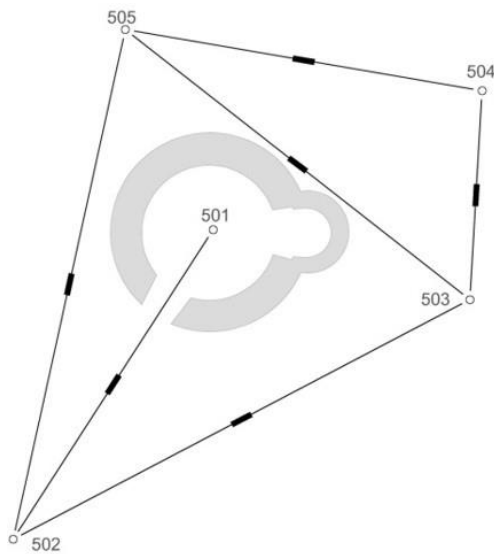


Figure 7. Free network around Kallósd church: sketch of control points, directions and distances

- 4.) The accurate measurement of the geodetic control point network. We need a continuous geodetic control point network based on direction and distance measurement both inside and outside the building (Fig. 7) and on more floors if necessary. Our aim is to provide a uniform, homogenous and accurate coordinate system. Accurate measurement means that we set up all the tripods with base plates, then we change the instrument and the prisms on the base plate to avoid positioning errors which can be dangerous due to the short distances (Fig. 8). As a result, we can ensure a network with deviations of one or two mm not only horizontally but also vertically. We must take a sufficient number of extra measurements (more than geometrically necessary).



Figure 8. Measuring control point network at Kallósd church: constrain centered setting up

- 5.) Precise measurement of detail points as polar points. Detail points (points for observation) are to

be measured at the same time as the control points using the same measuring equipment. Since they are chiefly building corner points, column corner points and arc points, where the positioning of a prism is not possible centrally, it is better to measure all these points without a prism. A card should be placed on the point to be measured so that it is perpendicular to the direction line and the touchpoint (line) of the card and the building must be set by the measuring equipment (Fig. 9). If a building point is covered by something, we need to employ a method that relies on points outside. Wall surface points perpendicular to the direction line can obviously be measured without a card. Clearly identifiable points are to be measured from two or more station positions.



Figure 9. Measuring wall-points without prism, using card

- 6.) Calculating the coordinates of the geodetic control point network and the points for observation. The calculation of the coordinates and the altitude of the geodetic control points must be carried out with adjustment or else we cannot take into consideration all the measurements simultaneously and, therefore, the result will not be accurate. The coordinate deviations cannot exceed 3-5 mm. Not only the coordinates of the control points, but also those points for observation must be shown with an accuracy of some mm. The calculation is to be done in a separate system as a free network to avoid frame errors that can influence the result. If the building has a standard axis, it is better to turn the local system using a planar isometry so that one coordinate axis should be perpendicular to the main axis of the building. We also apply a planar transformation when we wish to export our points to another system (to an adjacent one) – to illustrate how the axis of a church matches a cardinal point, for instance. For such transformation purposes the control points around the building ought to be measured by GNSS technology.
- 7.) Calculating the standard floor plan measures of the building. This calculation is carried out based on the measured points for observation using methods of coordinate geometry with an accuracy of mm. Measures of length and width, wall thickness, sizes of columns, distances of column intervals and altitudes belong here. Providing the standard data of the circles in terms of round churches is considered to be a separate task. An adjusted (regression) circle must be fit to the measured points of the arcs according to the least squares method (Fig. 10). We will obtain not only the

coordinates of the centre (C) and the radius (R) of the circle but also the standard deviations (rms) errors of them. Depending on the residuals (v) and deviations (rms) we are able to decide whether the building suits our purposes or not. Furthermore, standard deviations can be useful when determining the weight of measures later on.

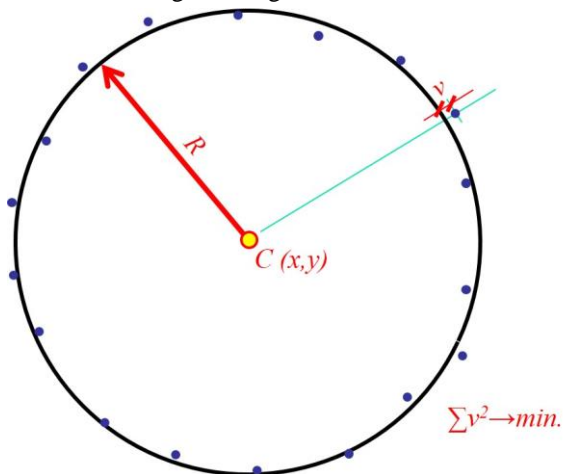


Figure 10. Symbolizing the adjusted circle

- 8.) Constructing the floor plan. The floor plan of the building is now ready to be constructed based on the points that have been measured and the calculated measurement data – in a metric system, of course. The points (in the same vertical plane) that are on one line can be drawn as a regression line. Standard measures are given with an accuracy of some mm.
- 9.) Matching standard building measurements to the former units of length. It is best to do this in an Excel table. The first questions we wish to answer is: Was the unit of length at the time of the construction expressed as royal foot or royal span? We get the desired information after dividing the standard, metric building measurements by the 'official' metric length of the foot (0.3126 m) and that of the span (0.1954 m). For some units of length we get round (or half) pieces which are regarded as preliminary values.
- 10.) Constructing the floor plan using the former unit of length. We try to make a floor plan that is similar to what the original could have been where standard measures were probably given in round (or half) multiples of the foot or span. It is really time-consuming and we may not succeed at once.
- 11.) Reconstructing the former unit of length. We need to make a table which includes both the standard distance data of the building given in metres and the unit of length in pieces. The metric value of the former unit, usually expressed as cm, is obtained from the quotient of the two data. The metric values of the unit will certainly not be the same. We recommend such a weighted average as an end result where we take into consideration how well the two endpoints of the measure observed were identifiable. For example, the deviations of the radii are helpful in this case.

The technology described above has been invented after surveying several buildings and processing their measures.

We can highly recommend it for similar purposes. In our view, appropriate results can only be obtained using accurate surveying methods. This is how surveying contributes to reconstructing the length etalon.

IV. THE RECONSTRUCTION OF THE UNIT OF LENGTH BASED ON THE ACCURATE SIZE DETERMINATION OF FIVE ROUND CHURCHES

In this chapter we present our practical surveying work and some results of it. The workflow of surveying method was the same we mentioned in Chapter III. Five Hungarian round churches were surveyed and analysed to reconstruct the ancient unit of length

A. The Saint Anne round church in Kallósd

Kallósd is a small bag village in Zala county. Its parish church was built around 1270 in Romanesque style.



Figure 11. Round church of Kallósd

The inhabitants of the village were forced to leave the church by reason of the Turkish occupation in the 17th century. The population returned in 1711. They started to clean the thicket around the church and renovated the building in 1740. Because of the growing number of inhabitants they had to build a hallway to the church in the 19th century which was demolished during the renovation between 1989 and 1993 in order to preserve the round church in its original shape.



Figure 12. Specialities of Kallósd: sitting bays and lizenas

The walls are built from brick. it means the identification of walls and measuring points are ideal for our purposes. There are seven sitting bays (sedilia) inside the nave, the sizes of them are also interesting for us. The other specialities are the small columns outside the nave wall a so called lizenas. 9 of them are on the northern part (left to entrance) and 3 of them are on the southern part (right to entrance). The results are seen in the Table II.

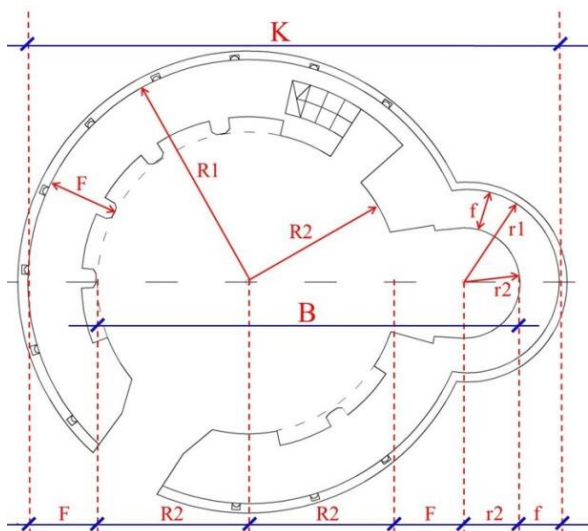


Figure 13. Floor plan and notations

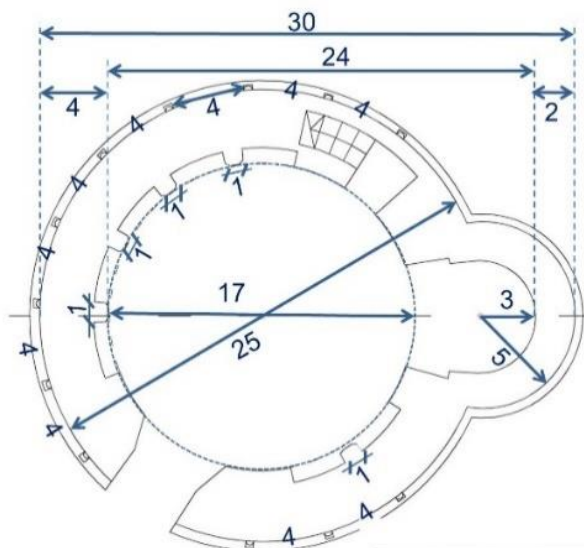


Figure 14. Floor plan of Kallósd in former unit (in foot)

TABLE II.
THE SIZES OF KALLÓSD CHURCH

	Description of sizes		Distance (meter)	RMS (meter)	Pieces.	foot (cm)	weight
1	Inner radius of nave (from 20 points)	R2	2,671	0,003	8,5	31,42	3
2	Outer radius of nave (from 20 points)	R1	3,937	0,003	12,5	31,50	3
5	Inner radius of sanctuary (from 6 points)	r2	0,980	0,012	3	32,67	1
4	Outer radius of sanctuary (from 8 points)	r1	1,627	0,006	5	32,54	2
3	Outer length of the church ($2R1+r1$)	K	9,501	0,008	30	31,67	2
6	Inner length of the church ($2R2+F+r2$)	B	7,588	0,019	24	31,62	2
7	Thickness of nave wall ($R1-R2$)	F	1,266	0,005	4	31,65	2
8	Thickness of sanctuary wall ($r1-r2$)	f	0,647	0,014	2	32,35	1
9	Lizena width (12)		0,155	0,002	0,5	31,00	1
10	Distance between lizenas (10)		1,267	0,003	4	31,68	1
11	Column width at sitting bays (5)	c	0,314	0,002	1	31,40	1
12	Radial size of columns (4×8)	a, b	0,204	0,002	0,625	32,64	1
13	Width of sitting bays (7)		0,99	0,004	3,125	31,68	1
14	Radial size of lizenas (2×12)		0,101	0,004	0,3125	32,32	1

B. The round church of Bagod



Figure 15. Bagod church outside and inside today

Bagod is a settlement in Zala county, earlier on its territory 3 independent villages were found: Bagod, Vitenyéd and Szentpál. Nowadays Szentpál is a bag village, in the cemetery of this village situated the earliest rotunda. It was built at the end of 13th century, but in 18th century it was totally rebuilt and enlarged. The original nave became as a sanctuary and new rectangular nave was built.

During 20th century the church was deserted, the roof was destroyed. Between 1999 and 2001 the church was totally renovated.

From our point of view only the today's sanctuary (the original nave) is interesting. There are four circles which could measure well: the inner and outer wall, the foundation and the circle of sitting bays. The results are seen in the Table III.

TABLE III.
THE SIZES OF BAGOD CHURCH

	Description of sizes		Distance (meter)	RMS (meter)	Pieces.	foot (cm)	weight
1	Outer radius of original sanctuary (from 5 points)	r1	2,590	0,001	8	32,38	0,5
2	Inner radius of original sanctuary (from 8 points)	r2	1,784	0,002	5,5	32,44	0,5
5	Outer radius of original nave (from 16 points)	R1	4,137	0,001	13	31,82	2
4	Radius of nave foundation (from 24 points)	R3	4,312	0,002	13,5	31,94	2
3	Inner radius of nave (from 13 points)	R2	2,566	0,001	8	32,08	2
6	Radius of sitting bays (from 6 points)	R4	2,880	0,061	9	32,00	1
7	Thickness of nave wall	F	1,571	0,001	5	31,42	1
8	Thickness of sanctuary wall	f	0,806	0,002	2,5	32,24	1
9	Total outer length of original church	K	10,171	from floor plan	32	31,78	0,5
10	Inner length of the original church	B	7,790		24,5	31,80	0,5
11	Lizena width		0,485	0,004	1,5	32,33	1
12	Radial size of sitting bays		0,314	0,061	1	31,40	1

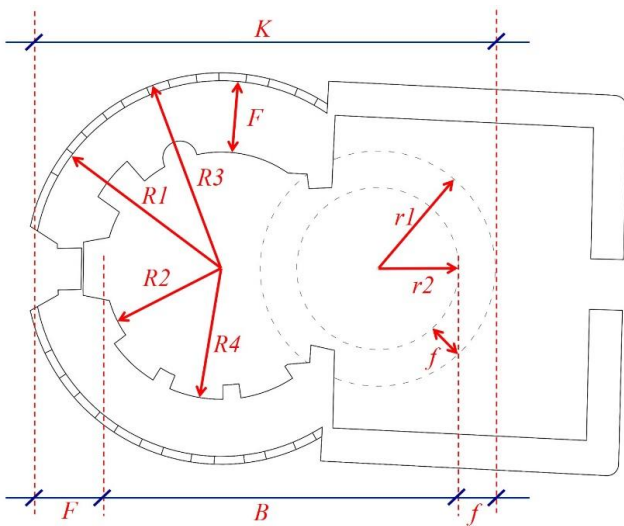


Figure 16. The floor plan and notations of Bagod church

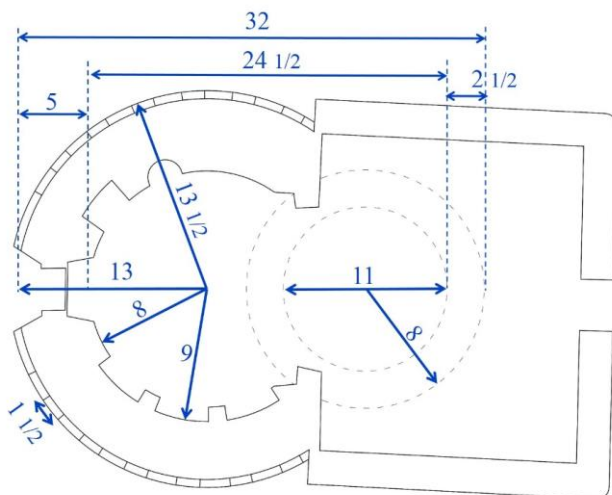


Figure 17. The floor plan in original unit (in foot)ch

C. *The Saint James church of Ják with four vaults*



Figure 18. The St James round chapel (left) and the Benedictian monastery church in Ják (foto:Civertan)

Two churches were built in the middle of 13th century in Ják settlement (Vas county): one as benedictian monastery church and one as presbyterian chapel. The first one is the famous monument of Hungarian medieval architecture, the second one is the small Saint James round chapel with four vaults, serving as a church of the village in the Middle Ages.

As it has been proven by the excavation in 1997 the St James chapel have been built on a rotunda foundation also. The curves of this earliest rotunda is now seen on the brick floor.

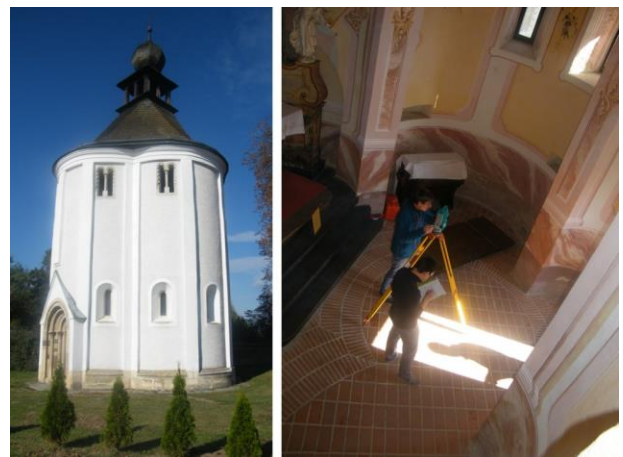


Figure 19. Ják church outside and measuring inside

TABLE IV.
COORDINATES OF CIRCLE CENTERS, RADIUS AND ITS RMS OF JÁK CHURCH WITH FOUR VAULTS

Description	Centre	y	x	Radius	r	RMS y	RMS x	RMS r
1 st arc, inner wall	K1 (Northern arc)	499,850	202,025	r1	1,492	0.010	0.019	0.009
1 st arc, outer wall		499,852	202,009	R1	2,620	0.002	0.007	0.002
1 st arc, foundation		499,848	202,014	RL1	2,815	0.002	0.005	0.001
2 nd arc, inner wall	K2 (Eastern arc)	502,046	200,183	r2	1,501	0.012	0.002	0.005
2 nd arc, outer wall		502,049	200,210	R2	2,603	0.007	0.002	0.002
2 nd arc, foundation		502,076	200,217	RL2	2,799	0.011	0.004	0.002
3 rd arc, inner wall	K3 (Southern arc)	500,208	198,014	r3	1,487	0.003	0.007	0.003
3 rd arc, outer wall		500,197	198,015	R3	2,607	0.003	0.009	0.003
3 rd arc, foundation		500,219	197,988	RL3	2,795	0.005	0.011	0.002
4 th arc, inner wall	K4 (Western arc)	498,078	199,844	r4	1,515	0.032	0.008	0.030
4 th arc, outer wall		498,050	199,843	R4	2,608	0.035	0.006	0.012
4 th arc, foundation		498,055	199,861	RL4	2,803	0.027	0.007	0.005

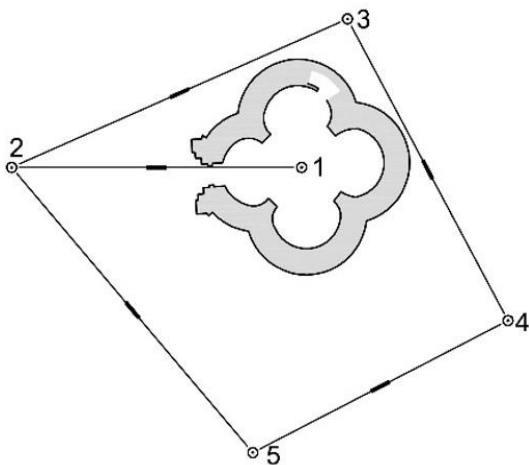


Figure 20. The sketch of control point micronet around the church

We set up a micro-net around the chapel with 5 control points (Fig. 20) and measured all detail points (in every meter sequentially). We identified 3 circles (arcs) in each vault: the inner and outer wall points and the outer foundation points (12 arcs altogether). After it we calculated the center point coordinates and radiuses of these arcs (Table IV). The standard deviations of these parameters are below 1 centimetre only one exception is the western arc. The reason is that on this vault we could measure few points because of the entrance door.

The centres (K1, K2, K3, K4) of different adjusted circles are mainly the same. These centres are corner points of square. The interesting thing is that K1-K4 centres and the wall endpoints (F1-F4) are located on the same circle.

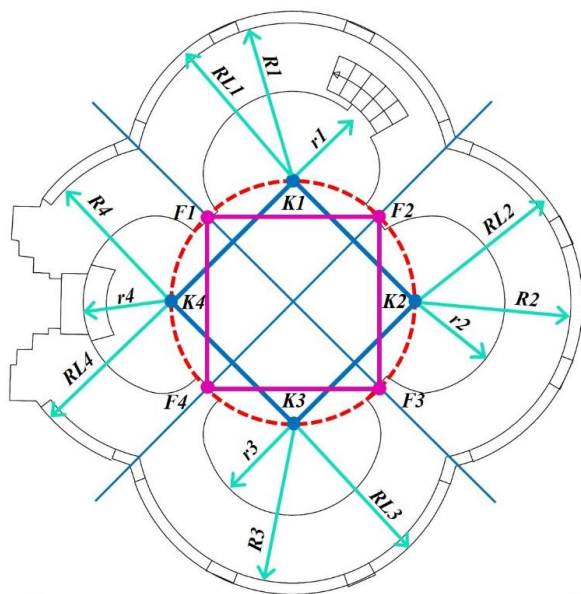


Figure 21. The ground plan of Ják church and notations

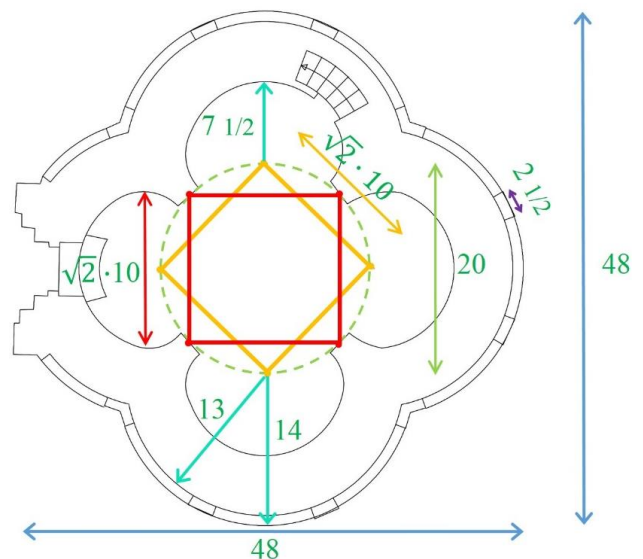


Figure 22. The sizes of Ják chapel in old Hungarian spans

What is the radius size of this circle? If we analyse and examine the size we will find that the size of radius is exactly 10 span. So it means that the sizes of tis chapel are not determined on foot but in span. All other radiuses we can determine in integral number of spans, for example

TABLE V.
 THE SIZES OF JÁK CHAPEL

	Description of sizes		Distance (meter)	RMS (meter)	Pieces.	span (cm)	weight
1	1 st arc,) radius of inner wall (from 8 points)	r1	1,492	0,009	7,5	19,89	1
2	1 st arc, radius of outer wall (from 12 points)	R1	2,620	0,002	13	20,15	2
5	1 st arc, radius of foundation (from 16 points)	RL1	2,815	0,001	14	20,11	2
4	2 nd arc, radius of inner wall (from 6 points)	r2	1,501	0,005	7,5	20,01	1
3	2 nd arc, radius of outer wall (from 16 points)	R2	2,603	0,002	13	20,02	2
6	2 nd arc, radius of foundation (from 16 points)	RL2	2,799	0,002	14	19,99	2
7	3 rd arc, radius of inner wall (from 6 points)	r3	1,487	0,003	7,5	19,83	1
8	3 rd arc, radius of outer wall (from 13 points)	R3	2,607	0,003	13	20,05	2
9	3 rd arc, radius of foundation (from 17 points)	RL3	2,795	0,002	14	19,96	2
10	4 st arc, radius of inner wall (from 6 points)	r4	1,515	0,030	7,5	20,20	1
11	4 st arc, radius of outer wall (from 7 points)	R4	2,608	0,012	13	20,06	2
12	4 st arc, radius of foundation (from 11 points)	RL4	2,803	0,005	14	20,02	2
13	Distance between K1-K2 points		2,856	0,007	14,14	20,19	1
14	Distance between K2-K3 points		2,872	0,005	14,14	20,31	1
15	Distance between K3-K4 points		2,830	0,014	14,14	20,01	1
16	Distance between K4-K1 points		2,810	0,013	14,14	19,87	1
17	Distance between F1-F2 points		2,822		14,14	19,96	0,5
18	Distance between F2-F3 points		2,814		14,14	19,90	0,5
19	Distance between F3-F4 points		2,799		14,14	19,79	0,5
20	Distance between F4-F1 points		2,822		14,14	19,96	0,5
21	Total outer length (E-W)		9,613		48	20,03	2
22	Total outer length (S-N)		9,633		48	20,07	2
23	lizen width (6 pieces)		0,506	0,013	2,5	20,24	1

the radius of foundation arc is 14 span, the radius of outer wall is 13 span, the radius of inner wall is 7 and half span.

The other interesting thing is that the total length of this chapel is 48 span it means exactly 3 old Hungarian fathom (Table V).

V. CONCLUSIONS

 TABLE VI.
 THE RECONSTRUCTED HUNGARIAN FOOT UNIT
 FROM ALL SIZES OF THREE BUILDINGS

	span	foot (cm)
Kallósd church		31,78
Bagod church		31,92
Ják chapel	20,03→	32,04
mean:		31,91

We nicely measured the sizes of three round churches from Medieval Ages, redrew the floor plan of these buildings. The measures were first given in metres but later in the ancient unit of length. These floor plans were used to recalculate the size of the royal foot in metres. We

used all the sizes and calculated the weighted average for all three buildings (Table VÍ).

At the end, as the average size of ancient Hungarian royal foot we got 31.91 centimetres. So we reconstruct the earlier original unit of length in a metric system. It means that the Hungarian royal fathom (10 feet) equals 3.19 metres instead of the 'official' value of 3.126 metres. Our assumption and the reconstruction method to obtain conversion factors were thus confirmed.

REFERENCES

- [1] Bogdán István: Magyarországi hossz- és földmértékek a XVI. század végéig. (Length and squares measures in Hungary until the end of XVI. century). *Akadémiai Kiadó, Budapest*, 1978. p 388.
- [2] Alán Kralovánszky: The Earliest Church of Alba Civitas. *Alba Regia, Székesfehérvár*, 1983, Vol. XX. pp. 75-88.
- [3] Busics Gy.-Tóth S.-Páli M.: Az egykori hosszegység meghatározása két megmaradt középkori templom méretei alapján. (The determination of ancient length unit of measurement by the sizes of two churches from medieval ages). *Geodézia és Kartográfia, 2016/3-4*. pp. 7-12.

Use Python in the Information Technology Practice Exam

László Gugolya

* OE AMK, Székesfehérvár, Hungary
gugolya.laszlo@amk.uni-obuda.hu

Abstract— Since 2005 new graduation system in Hungary. In this context, the advanced level practical exam should also be subject to programming. Here you can choose between the candidates of the programming languages. The optional languages later (2012) was added to the Python language. This language is widely used for educational purposes in the US universities. The language is spreading in Hungary, that look good to the baccalaureate exams students at elementary and secondary level elevated nearly 10% of this. More and more people are chosen from among the teachers in the classroom and students ' exams.

This trend has been placed under investigation, following which the other options for the language in the language compared to the baccalaureate curriculum. What are the advantages and disadvantages are to be expected in the course of teaching and learning.

I. INTRODUCTION

In the current maturity system, the candidate can choose from several programming languages. Thus, the question arises as to which language to choose the candidate and the teacher preparing the exam. There are several aspects to consider here. In the case of individual preparation, the candidate intends to continue to study as an important aspect. If you are going to learn Visual Basic in the higher education institution (eg economic informatics), then it is best to choose this option. This is not possible in a school or group environment. In this case, other considerations arise, such as the prior knowledge of the teacher and the usable time frame. Over the last few years, the number of hours spent on computing and within programming has decreased. Consequently, teachers are seeking ways to search for opportunities.

At present (2017), maturity exams can be selected from Pascal, C ++, C #, Visual Basic, Java, Python languages [1]. According to the programming language topology (TIOBE index) [2], this is a list of trends, from which Pascal hangs a bit from the line. This programming language has historically played a significant role in the Hungarian computer science education and is currently being implemented. So it can be justified to include it on the list.

I have been preparing for graduation for many years, but so far Python has not been educated. Several colleagues have a positive experience of using and teaching the language. So I came to see the time to look more closely at the potential. If I get caught, I tried to read on the Internet or read it in the literature [4].

Aug 2017	Aug 2016	Change	Programming Language	Percentage	Change
1	1	↓	Java	17.561%	-0.00%
2	2	↓	C	6.471%	-0.00%
3	3	↓	C++	5.020%	-0.02%
4	4	↓	C#	4.100%	-0.11%
5	5	↓	Python	3.880%	+0.31%
6	6	↓	Visual Basic .NET	2.390%	+0.00%
7	7	↓	PHP	2.070%	-0.00%
8	8	↓	JavaScript	2.000%	-0.01%
9	9	↓	Perl	1.880%	-0.02%
10	10	↓	Go	1.860%	+0.11%
11	11	↓	SQL	1.220%	-0.00%
12	12	↓	Object Pascal (Delphi)	1.000%	-0.00%
13	13	↓	Visual Basic	1.000%	-0.00%

Figure 1. TIOBE lista, 2017 augusztus

A. Dating

After the first chat, reading the Internet, you can place the language you want to know. Python is a general purpose, high level programming language. When designing the language, the readability and facilitation of programming work were emphasized.

Python supports functional, object-oriented, imperative, and procedural programming paradigms. It uses dynamic types and automatic memory management. The Python interpreter language, ie the source and object code is not separated.[3]

We can use it in a variety of areas: for web applications, desktop applications, game development, but many systems use it as a snap language (SPSS, PostreSQL).

There are several tools for programming. You can choose between IDLE and WinPython in maturity exams. For the learning process, we can use any of our favorite text editor, as it is widely supported because of the prevalence of the language. Larger systems also support this language, so they can be used in Visual Studio, Eclipse, Netbeans, as well. There are some systems that prefer this language. Such are the many popular PyCharms and NINJA IDEs. For the first steps, I used Komodo Edit, then I asked for a 1-year educational license for PyCharm and started learning about language.

II. FIRST STEPS

A. Language elements

When I get acquainted with the language elements, I use the 2017 code of the general exam level exam. You can download the text of the task and its source file from <http://oktatas.hu>. The official solution was released this year in C #. During the initial steps, I tried to use simple solutions. An important aspect was the ability to teach in learning.

Using the Python Console is a useful tool at the beginning of getting started. Here, you can see the result

of our instructions, and any outputs. This is very useful during the first step of the educational process.

The first significant difference is the significance of formatting compared to the other languages used in the exam. Missing the instruction block with the characters. This can be done by formatting. This in the first place is an advantage, as educational experiences show that it is natural for students. In the case of other languages, the use of code templates has less problems in this field.

```
1. for m in tesztek:
2.     if beKod == m["kod"]:
3.         print(m["megoldas"])
4.         keresett = m
```

The counting cycle shown in the previous example forms a statement block. The statement block of the conditional statement consists of the "print" and the assignment instructions.

The assignment was the usual marking. It is possible to multiply it. This can be seen in the first line. Multiple Assignments can be used in another form. This is equivalent to a = 0; b = 1. This instruction allows you to exchange values for variables. This is shown in the third row.

```
1. a=b=c=2
2. a, b = 0, 1
3. a, b = b, a
```

Conditional statements do not include a multiple conditional statement. This if ... elif ... must be solved. When using the terms, it is possible to use the relationship of mathematics classes. Example of solving problem 6.

```
1. if k <= 5:
2.     pontszam += 3
3. elif 6 <= k <= 10:
4.     pontszam += 4
5. elif 11 <= k <= 13:
6.     pontszam += 5
7. elif k == 14:
8.     pontszam += 6
```

The language knows the conditional term. Its shape differs from the usual C-like language.

```
1. kisebb = a if a < b else b
```

For cycles, the back tester is not among the language elements. This can be replaced by the first time with the (while) test.

```
1. while True:
2.     utasítások
3.     if <kilépési feltétel>:
4.         break
5.     (további utasítások)
```

Here is another language for the break and continue statements. They only apply to one cycle.

The "for" cycle can be used extensively. Its operation is fundamentally different from the usual. It does not go through a series of numbers, but it does enter something. That is, it is not a "counting" cycle, but a "crawling" cycle. Here is no cycle variable in other languages. This is more like a foreach in C#. We can go through all the elements of a multitude (eg a list). For example, for task 5.

```
1. dbJo = 0
2. for m in tesztek:
3.     if helyes[s] == m["megoldas"][s]:
4.         dbJo += 1
```

You can also use cycles in numeric order using a function. The range () function generates an interval as a return value, and the resulting list goes through the for cycle. This way we can produce a cycle close to a traditional one. This could be used to map the solutions stored in the string by character in task 6. Here, since 14 questions were included in the test, the range () function produces a 0-13 interval.

```
1. for k in range(0, len(helyes)):
2.     if helyes[k] == m["megoldas"][k]:
3.         if k <= 5:
4.             pontszam += 3
5.         elif 6 <= k <= 10:
6.             pontszam += 4
7.         elif 11 <= k <= 13:
8.             pontszam += 5
9.         elif k == 14:
10.            pontszam += 6
```

Repetitive structures may have another branch. This is different from the other programming language that can be used for the exam. This branch will be executed if the cycle has run through the list (for case) or if the condition is fake (while). It will not be executed if the cycle is interrupted by the break command.

```
1. while <feltétel>:
2.     utasítások
3. else:
4.     utasítások
```

Many people like to solve each task individually. This will make the solution more understandable. To do this you need to know how to use the functions. The function is specified as follows.

```
1. def feladat2():
2.     print("2. feladat: ")
3.     print(" A vetélkedőn {0} versenyző
indult.".format(len(tesztek)))
```

You can specify parameters as usual, return the value to the return. Here's an interesting opportunity to see, this is multiple value reproduction. In the example below, 6 and 9 are printed.

```
1. def fuggveny(x):
2.     return x*2, x*3
3. a,b = fuggveny(3)
```

```
4. print (a, " ",b)
```

It is also possible to enter local functions similarly to the Pascal language.

```
1. def fgv():
2.     def alfgv():
3.         print("alfgv vagyok")
4.     print("fgv vagyok")
5.     alfgv()
```

The data structure like arrays is multi way in Python: list, tuple, dictionary.

The most versatile composite data type of Python is a list (list) that can be entered as comma separated values in square brackets. The elements in the list do not have to be of the same type.

```
1. gyumi=["alma",260, True, 12.5]
2. gyumi2=[["alma","körte","meggy"],260]
```

The second line shows that the list item can be a list. You can refer to the element of that list by block from 0. It is also possible to refer to several elements of the list (slices, parts), for example:

```
1. gyumi[1:3]->[260, True]
```

It is also possible to access the list from the back, so negative indices are used. For example, the last element can be referred to as a `gyumi[-1]`.

To manage lists, you can use `append`, `extend`, `insert`, `remove`, `pop`, `index`, `count`, `sort`, `reverse`. These can be used as stack or row data structures. This is useful for problem solving.

"Two-dimensional array" is used in maturity exams. We can do this with a list in the list. This is like a one-dimensional array implemented in one-dimensional array in other languages.

```
1. lista=[[2,3],[5,6],[8,9]]
2. s = ""
3. for x in lista:
4.     for y in x:
5.         s += " " + str(y)
6.     s += "\n"
7. print(s)
```

Listing is facilitated by so-called "list mapping". You can then perform a particular action on all the items on the list, and then create a new list. You can view it quickly on the console.

```
1. >>> lista=[1,2,3,4]
2. >>> lista = [elem*2 for elem in lista]
3. >>> lista
4. [2, 4, 6, 8]
```

Like a list, a structured data structure for storing different objects is tuple. Contrary to the list, the elements can not be modified here. This may be useful if you have

to work with fixed elements, such as the names of days and months.

```
1. napok = ('hétfő','kedd','szerda','csütö
rtök','péntek','szombat','vasárnap')
```

Use dictionary to perform record-like data storage. You can store key-value pairs here. As an example, the 2013 "Választás" taskbar, where you can enter pairs of party abbreviations and names. Using the dictionaries with a list, we can implement traditional, structured, record-keeping data storage.

```
1. partok={"GYEP":"Gyümölcsevők Pártja",
2.         "HEP":"Húsevők Pártja",
3.         "TISZ":"Tejivők Szövetsége",
4.         "ZEP":"Zöldségevők Pártja",
5.         "-":"Független jelöltek"}
```

The String object can not be modified. Its use is the same as in the lists, that is, a list of roundabouts. We have methods for implementing string actions. Of these, the use of the `strip` should be highlighted. After scanning (console, file), we need to remove whitespace characters. You can do this with the `strip()`.

Handling sets as a standalone data type can be implemented. Creating it with `set()`. This had to be repeatedly used in the maturity quiz of recent years. For example, in the May 2013 taskbar, to define themes.

```
1. temat=set()
2. for tema in adatok:
3.     temat.add(tema)
4. print("A tematörök:", ", ".join(temak))
```

You can create a set from an existing list.

```
1. >>> kosar = ['alma', 'meggy', 'alma', '
cseresznye', 'meggy', 'alma']
2. >>> gyumi = set(kosar)
3. >>> gyumi
4. {'meggy', 'alma', 'cseresznye'}
```

The usual actions can be done on the sets: containment, embedding, deleting, engraving, union, difference, symmetric difference (`in`, `add`, `pop`, `remove`, `discard`, `&`, `|`, `^`).

In tasks, it is often necessary to sort the stored data. We now have the option to use traditional sorting algorithms, but the language provides a way to sort the lists.

```
1. lista = [5, 3, 4, 1, 2]
2. for i in range(len(lista) - 1):
3.     for j in range(i + 1, len(lista)):
4.         if lista[i] > lista[j]:
5.             lista[i], lista[j] = lista[
j], lista[i]
6. print(lista)
7. lista = [5, 3, 4, 1, 2]
8. lista = sorted(lista)
9. print(lista)
10. lista = [5, 3, 4, 1, 2]
```

```
11. lista.sort(reverse=True)
12. print(lista)
```

In the example, the bubble algorithm is first seen. then the sorted () function, which creates a new list that is generated as a parameter. In the third solution, sorting is done locally. Sorting is done in descending order by specifying the parameter. For a more complex list, sorting () is sorted by the first "column". If you want something else, you can set this with the key parameter.

```
1. import collections
2. Diak = collections.namedtuple("Diak", "Nev,Szulido,Magassag,Tomeg")
3. peldaLista=[]
4. peldaLista.append(Diak("Nagy Anna", "2000-10-10", 170, 70))
5. peldaLista.append(Diak("Szép Éva", "2012-10-10", 120, 30))
6. peldaLista.append(Diak("Nagy Attila", "2011-1-21", 180, 68))
7. print(*sorted(peldaLista, key=lambda adat: adat.Tomeg), sep='\n')
8.
9. peldaLista2=[]
10. peldaLista2.append(["Nagy Anna", "2000-10-10", 170, 70])
11. peldaLista2.append(["Szép Éva", "2012-10-10", 120, 30])
12. peldaLista2.append(["Nagy Attila", "2011-1-21", 180, 68])
13. print(*sorted(peldaLista2, key=lambda adat: adat[3]), sep='\n')
```

In the first example, we see a solution when naming the column data stored in the list and naming it for the sort key. In the second example there is a list with a list to complete the sorting. You will then need to enter the "column" number at the key.

Reading like array:

```
1. fajl=open("valaszok.txt")
2. adat =[]
3. for e in fajl.readlines():
4.     adat.append(e.strip().split())
5. print(adat)
```

Reading like record in the dictionary:

```
1. adat2=[]
2. fajl=open("valaszok.txt")
3. i=1
4. for e in fajl.readlines()[1:-1]:
5.     (kod,megoldas) = e.strip().split()
6.     valasz = {
7.         "sorsz": i,
8.         "kod": kod,
9.         "megoldas": megoldas
10.    }
11.     adat2.append(valasz)
12.     i+=1
13. print(adat2)
```

In case of random number generation, we can choose from several solutions.

```
1. import random
2. print(random.randint(1, 6))
3. print(random.random() * 100)
4. print(random.choice(['alma', 'meggy', 'cseresznye']))
5. print(random.randrange(0, 101, 5))
```

The autumn maturity of 2015 had to produce random coins at random, so we can easily do that on the basis of the above.

```
1. import random
2. print("A pénzfeldobás eredménye:", random.choice("IF"))
```

In the 2016 "Zár" task sequence, a series of codes had to be produced. This is also easy to solve. With the sample method, you can create a given number of samples from a given pattern and then merge it. You can see it on a bracket.

```
1. sorozat = random.sample("0123456789", 5)
2. sorozat
3. ['0', '3', '5', '7', '8']
4. print("".join(sorozat))
5. 03578
```

B. Experience in solving task series

After the basic elements of the language I tested a complete set of tasks. I solved the last "Test Competition" task in 2017. This can be considered as a mixed task, as it is possible to use record-based thinking, and handling text data (strings) is also needed. The exact description of the task and the source file can be found at https://dari.oktatas.hu/kir/erettsegi/okev_doc/erettsegi_2017/e_inf_17maj_fl.pdf.

The task series evaluates a contestant's response to test tasks by using a text file. The tasks have been solved by means of subprograms for ease of comprehension.

The first task is to read the data. The text file has been included in a global list of data.

```
1. def feladat1():
2.     print("1. feladat:")
3.     adatok=[]
4.     fajl=open("valaszok.txt")
5.     global helyes
6.     helyes = fajl.readline().strip()
7.     #további adatok beolvasasa
8.     i = 1
9.     for sor in fajl:
10.        sor = sor.strip()
11.        reszek = sor.split()
12.        valasz = {
13.            "sorsz": i,
14.            "kod": reszek[0],
15.            "megoldas": reszek[1]
16.        }
```

```

17.         adatok.append(valasz)
18.         i = i+1
19.         #print(adatok)
20.         fajl.close()
21.         return adatok

```

The second task is for the starting competitor. This can be done by using the length of the list.

```

1. def feladat2():
2.     print("2. feladat: ")
3.     print(" A vetélkedőn {0} versenyző
indult.".format(len(tesztek)))

```

The third task is to request a competitor's details. Here is a linear search query.

```

1. def feladat3():
2.     print("3. feladat:")
3.     global keresett
4.     beKod = input("A versenyző azonosít
ója = ")
5.     for m in tesztek:
6.         if beKod == m["kod"] :
7.             print(m["megoldas"])
8.             keresett = m

```

In the fourth task, the correct solutions of the competitor requested in the given form have to be given. The "sought-after" contestant's solutions are per character and weigh the output into a text variable.

```

1. def feladat4():
2.     print("4. feladat")
3.     global helyes
4.     global keresett
5.     print("{0:s} (a helyes megoldás)".f
ormat(helyes))
6.     ki = ""
7.     for k in range(0,len(keresett["mego
ldas"])):
8.         if helyes[k]==keresett["mego
ldas"][k]:
9.             ki = ki + "+"
10.        else:
11.            ki = ki + " "
12.        print("{0:s} a versenyző helyes vál
aszai".format(ki))

```

In the fifth task, the success of a given task solution has to be expressed in percentage. When a solution is made, a count is to be made.

```

1. def feladat5():
2.     print("5. feladat")
3.     beFeladatSorszama = int(input("A fe
ladat sorszama = "))
4.     dbJo = 0
5.     for m in tesztek:
6.         if helyes[beFeladatSorszama] ==
m["megoldas"][beFeladatSorszama]:
7.             dbJo += 1

```

```

8.     print("A feladatra {0} fő, a versen
yzők {1}%-
a adott helyes választ.".format(dbJo,ro
und(dbJo/len(tesztek)*100,2))

```

In the sixth task, the score of the competitors must be saved to an output file. Scores are given in advance for each task. Two embedded cycles should be used in the solution. One of the contestants goes the other way for each competitor's responses. You can also use `elseif` for the scores. The solution described illustrates the standardization in mathematics, which is different from other programming languages. The list of points⁷ will be useful when solving the seventh task.

```

1. def feladat6():
2.     print("6. feladat:")
3.     global helyes
4.     global pontok7
5.     kiFajl = open("pontok.txt","w")
6.     for m in tesztek:
7.         pontszam = 0
8.         for k in range(0,len(helyes)):
9.             if helyes[k] == m["megoldas
"]][k]:
10.                if k <= 5:
11.                    pontszam += 3
12.                if 6 <= k <= 10:
13.                    pontszam += 4
14.                if 11 <= k <= 13:
15.                    pontszam += 5
16.                if k == 14:
17.                    pontszam += 6
18.                kiFajl.write("{0} {1}\n".format
(m["kod"],pontszam))
19.                pontok7.append([pontszam,m["kod
"]]
20.            kiFajl.close()

```

In the seventh task, prizes must be awarded based on the scores obtained. First of all, we perform the orderly scoring according to the scores, and then the prizes are paid out to ensure that the prize winners do not lose.

```

1. def feladat7():
2.     print("7. feladat:")
3.     global pontok7
4.     rendezett = sorted(pontok7, reverse
= True)
5.     db = 1
6.     i = 0
7.     while i<len(rendezett) and db <= 3:
8.         print(db, '.díj', '(' ,rendezett[i
][0], 'pont):', rendezett[i][1])
9.         if i + 1 < len(rendezett) and r
endezett[i+1][0] != rendezett[i][0]:
10.            db += 1
11.            i += 1

```

The task was not particularly difficult. Python could be solved in 45 minutes without practice.

III. SUMMARY

Compared to the solutions of Python's own and others, there is no significant difference between the official solutions and the comparison. Regarding solutions, there is no difference in the number of rows if we omit the blank of "{}" of the official solution. It is felt that data storage is more flexible than in other more traditional languages. In addition to the scans, the use of Python was convenient (and therefore more advantageous) for sorting.

During the study and testing of the language, it became clear to me why Python was growing fast. On the basis of the many positive feedback it can be stated that it is worth trying the language and introducing it into secondary

education. However, I consider it implausible that Python does not play a significant role in the Hungarian higher education. Thus, entering the higher education of the students who are going to take part can be disadvantageous.

REFERENCES

- [1] <https://www.tiobe.com/tiobe-index/>
- [2] https://www.oktatas.hu/koznevelas/erettségi/2017oszi_vizsgaidoszak/2017oszi_nyilvanos_anyagok_listaja
- [3] <http://nyelvek.inf.elte.hu/leirasok/Python/>
- [4] Mark Summerfield, "Python 3 programozás", Kiskapu Kiadó, 2009

Assessment of Basic Programming Knowledge for New Students of Informatics Engineering at Óbuda University

László Gugolya

* OE AMK, Székesfehérvár, Hungary
gugolya.laszlo@amk.uni-obuda.hu

Abstract— In the past years, higher education institutions have measured their knowledge of enrolled students. Then, in the knowledge of this, they had to catch up with catchy, leveling lessons for the poor performing students. The technical backgrounds usually measure mathematics and physics and measure incoming students. At the Alba Regia Technical College of Obuda University, the demand for IT was raised in the informatics fundamentals. On the basis of past experiences, students experience great differences in knowledge. So there is a need to assess the knowledge we have so far. Thus, in a fortunate situation, students with nearly the same preconceptions could be divided into common practice groups. This helps to educate and provide opportunities for talent management. In the thesis the evaluation of the questionnaire and the deduction of the experiences are presented.

I. INTRODUCTION

In AMK Óbuda University has been in the field of engineering for a number of years, including IT training. Over the years, several ideas and curricula have been formulated. At present, the introduction of curriculum E is currently being introduced, which includes the curricula 5. Continuous change is not surprising since IT is a constantly changing area. The University of Székesfehérvár has been operating as an independent center for several years. So you have greater autonomy, which allows you to take into account local specialties. To make the most of these opportunities, this year we conducted a questionnaire survey among incoming engineer-informatics students.

At the Óbuda University (similar to other universities), in-school students have been measuring mathematical and physical knowledge for many years. For these subjects, it is easy to determine the level they should know about the students, since it has centrally provided them. Thus, after the knowledge assessment, students are bound to catch up whenever necessary.

This is not so in terms of programming pre-requisites. Programming is taught in specialized vocational secondary schools / specialty schools, as part of the preparation for high school graduates in grammar schools. So there are significant differences between the incoming students. The situation is further hardened by the fact that many students are self-educated or otherwise self-serving. So you gain a lot of knowledge and pre-qualification.

In this year's first questionnaire we did not ask for material knowledge, but we were just wondering how much and what they were learning. What are your plans, what areas are interested in the coming students? We have not studied the depth of their knowledge. We want to realize this next year.

II. QUESTIONNAIRE

A. Design, measurement

The questionnaire was filled for the first time with newly arrived students. We did not want the questionnaire to be long and complicated first. Trusting that honest filling will be bigger. The questionnaire consists of 14 questions, 11 of which are professional.

The measurement was completed by the first semester of the software design and development course. A total of 95 people completed the questionnaire.

The results of the questionnaire were supported by SPSS [1] and MS Excel software.

B. Questions

At the beginning of the questionnaire there is a brief introduction, and he asks for the respondent's form of training, gender for maturity and school type of school.

Next, information and programming issues arise. First of all, the questionnaire was asked about the preliminary knowledge: whether it had matured from IT, whether OKJ training, what programming languages were learned within and outside the education system? Then the questions that students will study are their IT interests, orientation and students' future plans: whether they want to attend B.Sc., M.Sc., what kind of specialization they choose, what areas they are interested in programming, what kind of attitude (engineering or informatics), what kind of programming language would you learn? After completing the questionnaire, questions are asked about catching up and talent management.

III. EVALUATION

A. Analyzes

The questionnaire was completed by 95 first students, 54 of them B.Sc., 33 Fsz and 8 did not respond. The gender distribution is 82 women and 13 men. 52 high school graduates and 43 secondary vocational schools. This is foresaw, as grammar schools have a minimum of computing hours. Based on the maturity year, it is possible to deduce age distribution.

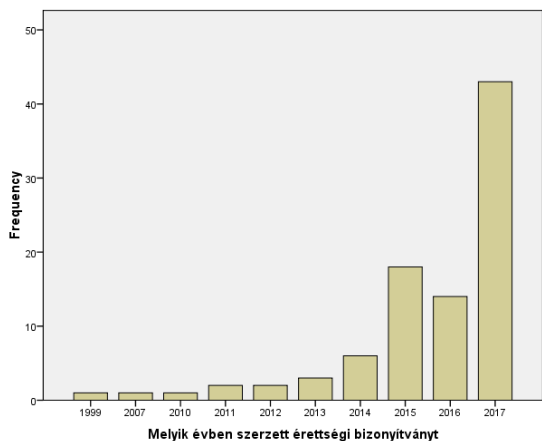


Figure 1. Question 4

Figure 1 shows that 79% of applicants have matured in the last 3 years.

The proportion of computer science graduates is shown in the following table (Figure 2).

Informatika érettségi vizsga típusa					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<i>nem</i>	10	10,5	11,5	11,5
	<i>közismereti középszintű</i>	62	65,3	71,3	82,8
	<i>közismereti emeltszintű</i>	8	8,4	9,2	92,0
	<i>ágazati középszintű</i>	5	5,3	5,7	97,7
	<i>ágazati emeltszintű</i>	2	2,1	2,3	100,0
	Total		87	91,6	100,0
Missing	System	8	8,4		
	Total	95	100,0		

Figure 2. Question 5

It can be seen from the table that a total of 15 students from the 95 year-old students selected a degree program where the programming skills are displayed. Based on this, it would be difficult to think of a group break that helps talent management. You have to choose another form for the university.

The following question revealed that 14 people had an ITC. This is surprising given the high school graduates, as they gain higher knowledge of OKJ training.

We also received unexpected values for the distant plans of students. 46% of respondents are planning to get a M.Sc. degree. This is not in line with the opinion of the graduate students.

An important area is exploring prior knowledge. Thus, two questions related to the programming languages learned so far. Since many people are learning to program outside of the education system, so there was a separate question.

Measured results for languages learned within the educational system. (Figure 3)

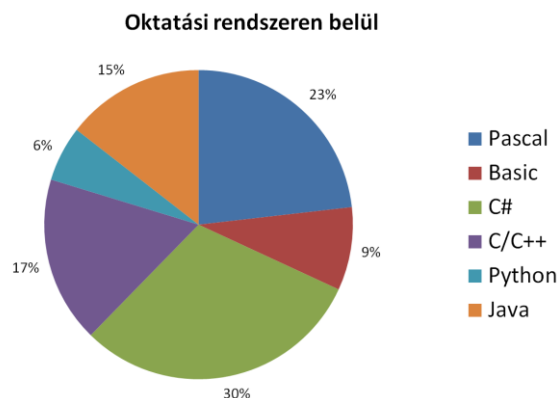


Figure 3. Question 4

Measured results for languages studied outside the educational system (Figure 4)

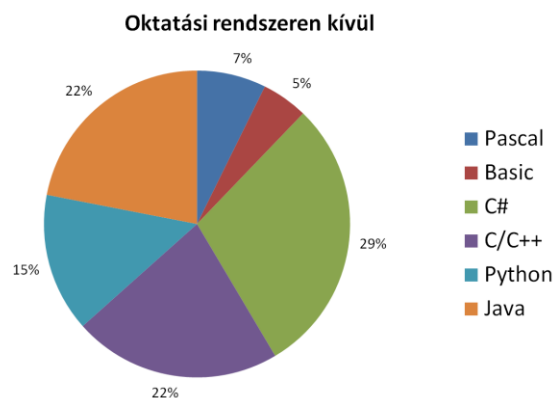


Figure 4. Question 4

Comparative summaries of the answers to the learned programming languages are shown in the figure below with the value table (Figure 5).

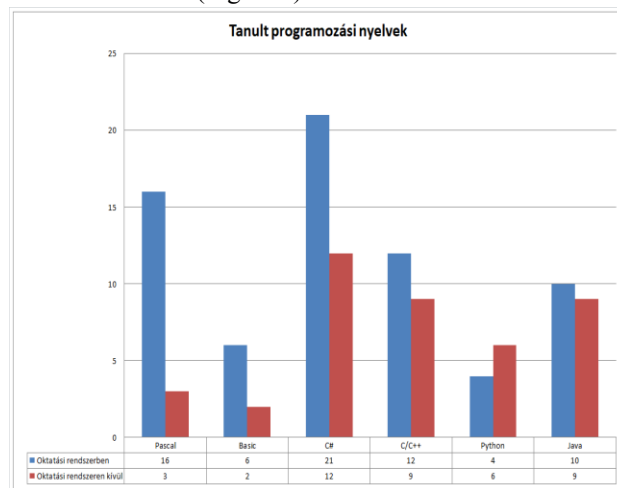


Figure 5. Question 8-9

In the course of the responses several students could be given the language. 41 enrolled within the educational

system programming language, languages, and 21 learned in other ways. Within the education system, Pascal is still a major player. Pascal was replaced by C#. In addition to the education system, the distribution is even more balanced for students. We can see a large number of Java and C/C++ students. Interestingly, Python is higher than the educational system value. This language is becoming more and more popular among both students and teachers, but it is not yet reachable in foreign countries. The main language of our university is C#. Based on the above data, this is a lucky choice. It is interesting to note that, despite the gradual rise of C# language, most universities do not follow this trend, as the University of Óbuda is the only one who uses this programming language as a basic language. More than the other response options, PHP and HTML have been typed. Responding to HTML refers to conceptual deficiencies, as this cannot be considered a programming language.

One central issue was the professional orientation of the students. The question was asked how students feel more affinity for the engineering course or the IT field?

A mérnök informatikus képzés tekintetében melyik irány áll közelebb hozzád?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid mérnök	8	8,4	8,4	8,4
informatikus	59	62,1	62,1	70,5
mindkettő	28	29,5	29,5	100,0
Total	95	100,0	100,0	

Figure 6. Question 10

The ratios are well illustrated by a pie chart from the table (Figure 6). (Figure 7)

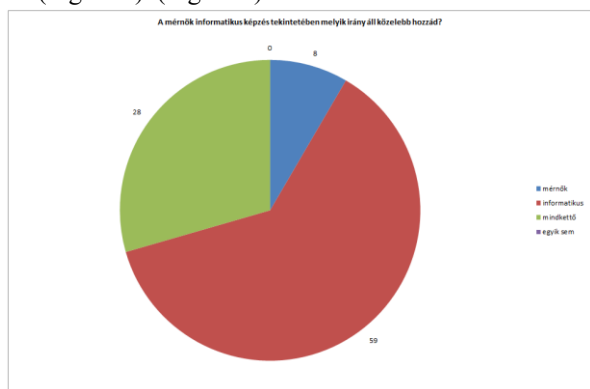


Figure 7. Question 10

It can be stated that students coming to the Alba Regia Technical Faculty are mostly interested in informatics skills. This is an important achievement, since when designing the specializations and the nature of the training, this must be taken into account. This is also reflected in the thesis work as a major part of the topics comes from software development.

Respondents could name the specialty they would like to deal with most. Here, the largest number of specialization related to network and software development has been named.

Then specific nominated areas of interest were selected by the students. The choices are:

- console applications / theory

- mobile applications
- web applications
- Robot programming
- desktop applications
- programming embedded systems
- games
- PLC programming
- database programming
- Other

The results are shown in the diagram (Figure 8).

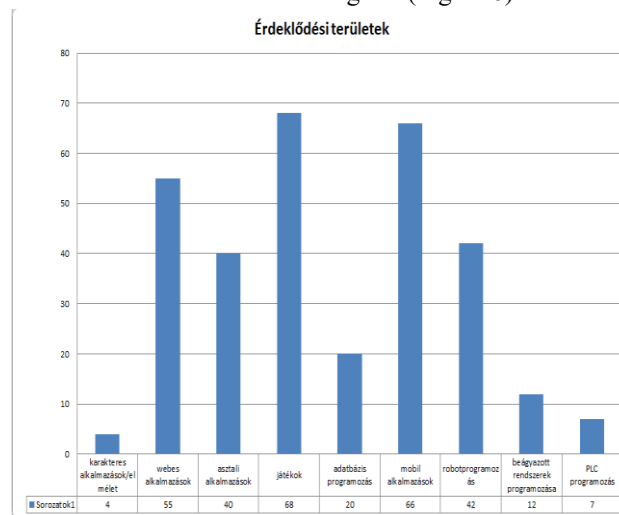


Figure 8. Question 12

The answers reflect current trends: games, mobile, web, and robots. (Figure 9).



Figure 9. Question 12

Then the respondents had to name it. Again, the same are the same among the answers.

The bar graph shows the relationship between listening students and talent management. It is interesting to note how much greater willingness to catch up than talent. (Figure 10.)

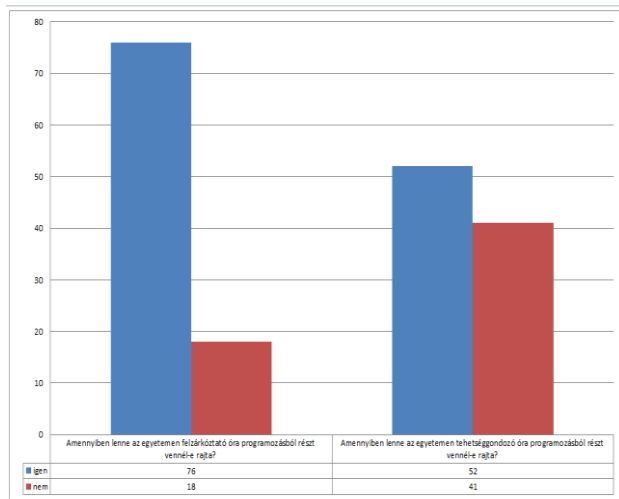


Figure 10. Question 12-13

Lastly, it was a question of how to learn if the student could choose. (Figure 11.)

1-Ha Te választanál milyen programozási nyelvet/nyelveket tanulnál az egyetemen?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	52,6	52,6	52,6
C	1	1,1	1,1	53,7
C#	19	20,0	20,0	73,7
C+	1	1,1	1,1	74,7
C++	7	7,4	7,4	82,1
C++ / C#	1	1,1	1,1	83,2
Java	8	8,4	8,4	91,6
Javascript	1	1,1	1,1	92,6
PHP	1	1,1	1,1	93,7
Python	6	6,3	6,3	100,0
Total	95	100,0	100,0	

Figure 11. Question 14

The frequency table shows the preferred C # language. In addition, C / C ++, Java, Python is popular among students. With respect to the curriculum [2], C # and Java are included in the master material, while C / C ++ can be learned as optional subjects.

IV. SUMMARY

The results basically proved the experience so far. Since our university has a good relationship between students and students, there were preliminary information on the subject. If we want to follow student expectations, the information technology of our training needs to be strengthened. Surveying the possibility of strengthening the training of software development. Network knowledge is currently a specialist. So this serves the student's needs.

Preliminary knowledge is strongly different from the questionnaire. Thus, in the coming years, it would be advisable to assess this quality, thus enabling opportunities for catching up and talent-providing.

REFERENCES

- [1] Sajtos László and Mitev Ariel: SPSS Kutatási és adatelemzési kézikönyv, Alinea Kiadó, 2007
- [2] <http://amk.uni-obuda.hu/index.php/hu/felvetelizoknek/21-mernok-informatikus>

Development of a Local Transportation Route Planner Android Mobile Application with Graph Algorithm

K. Zsobrak*, L. Gugolya*

* Óbuda University – Alba Regia Technical Faculty, Székesfehérvár, Hungary
 zsobrak.krisztian@gmail.com
 gugolya.laszlo@amk.uni-obuda.hu

Abstract—In this paper I describe how the Local Transportation Route Planner Android Mobile Application of Székesfehérvár was developed. The graph build and search algorithm was written in Java language and it was already ready to use, so the task was to create forms for the user data input and display the results in a proper way. Optimizations needed to be done on the graph build and search algorithm, due to the different resources on the mobile platform. Finally, the GUI was improved to be faster and nicer.

I. INTRODUCTION

Local transportation route planning applications of Székesfehérvár is accessible through the company’s web page, which is not suitable for mobile devices. Because this platform is very handy during travel, I decided to create the application for mobile devices.

At this point of the development, the graph build and search algorithm was already ready to use. It’s input parameters are the start and arrival station, and the start time. Its output is a list of possible travel routes. The data source is a database with the local transportation’s schedule. You may read about these topics in my previous papers.

II. PLATFORM

To choose the proper mobile platform I considered the distribution of the operating system in Hungary, and the conditions to develop and distribute the software. Currently there are three major operating systems on the market: Android, iOS, Windows.

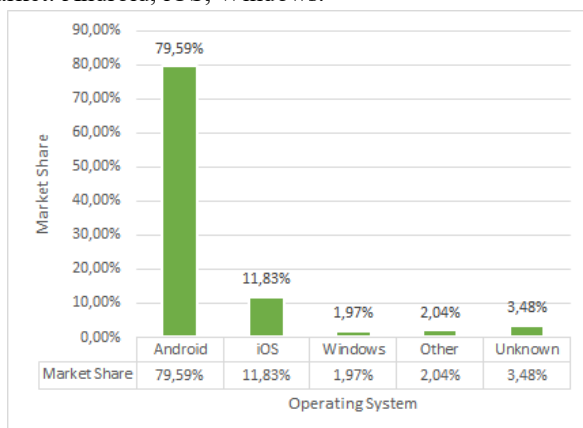


Figure 1. Market share of different mobile operating systems, 2017 September, Hungary [1]

The data shows that nearly 4 of 5 phones are using Android operating system in Hungary. Besides it is well distributed, the Android development software - the Android Studio - is free for use, thus it is the best choice.

The different versions of the operating system use different API levels, which are backward compatible. Application for the newer systems have more functionality, but have less devices to run on.

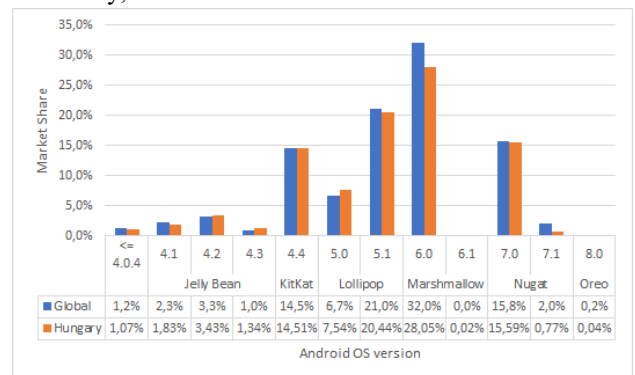


Figure 2. Market share of different Android versions, 2017 September, Hungary and worldwide [2][3]

The data on Figure 2. presents that the Hungarian and worldwide trends have correlation, but the newer systems are slightly less present here. Because the functionality required for this application are present in the early versions of the Android operating system, I decided to use the 4.1 version.

III. DESIGN

A. Database

The application does not require other databases than the schedule’s, because it is not using large amounts of data which would make it necessary. The design of the schedule’s database is described in my previous paper.

B. User Interface

On Android, the forms and data are displayed on Activities. This application emphasizes functionality, displaying each screen on different activity is the easiest.

The Main Activity is the starting activity. I will display quick information on this screen, now it displas a simple welcome message. Because the user can navigate to five different activities from here, the best solution was to use

a dropdown menu, which is placed on the top right corner. The most important activities have an information bubble on the bottom right corner, which displays a quick help about the form or the data on the screen.

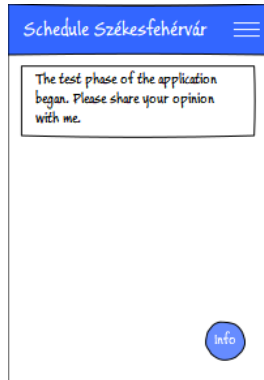


Figure 3. Main Activity plan

The route planning algorithm requires the start and the end station from the user. The best choice is using dropdown list to select one of Székesfehérvár's 171 stations. Often the passengers would like to travel back to the station they came from, it is practical to place a button to the UI which switches the start and the end station.

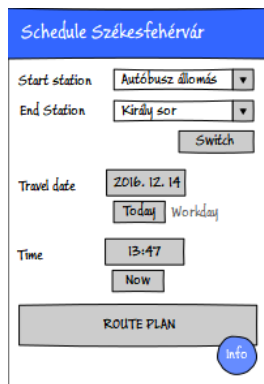


Figure 4. Route plan activity plan

To use the proper schedule, the application needs to know which date the user would like to travel, therefore, a date select option is required. The Android framework provides a date picker fragment, which can be displayed in a pop up window. To spare space on the form, I decided to display the selected date on the button. The default date is the actual today. If the user changes the date, and would like to go back to the actual day, pressing the "Today" labelled button will do it. Also, the software displays the selected day's work schedule (working day, non-working day, etc.) with a light gray color. This way the users can be sure, they selected the proper date.

The search algorithm also requires the start time, when the user plans to travel. To choose time, the Android framework provides a time picker fragment which could be displayed in a popup box too. As the date picker button, this also displays the selected time as the label of the button. There is also a button, labelled "Now", which changes the selected time to the default value.

After pressing the route planning button, the graph build and search algorithm starts to run. The results are displayed on another activity. Because this process will take a second at least, it is necessary to run this task on a

separate thread, otherwise the application would be non-responsive. Meanwhile a popup window displays an information about the process.

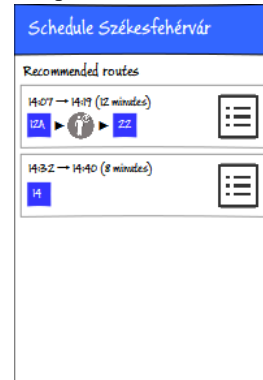


Figure 5. Route plan results Activity plan

The results of the search algorithm are stored in a list. The application presents this in an easily understandable way, where the start time, the arrival time, the length of the travel, and the required lines' number are displayed in a box. The screen must be scrollable, because the list may not fit on the device's screen in one piece. If the user clicks on a plan, the application switches to the next activity, where the details are shown.

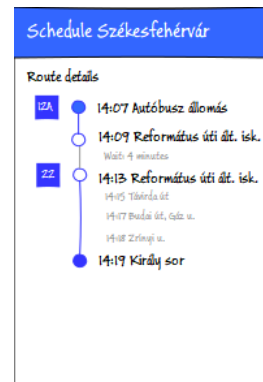


Figure 6. Route details Activity plan

On this activity, the user gets information about when the bus will arrive to the station, where he should change, and how long must he wait.

These "time data" is only predicted, because a few minutes delay from the schedule could be caused by the traffic.

IV. DEVELOPMENT

A. Convert MySQL database to SQLite

The best format to store the schedule's database on the device is SQLite [4], because the Android system handles it natively [5], a third-party software is not necessary, and it is easy to make queries through the Android API. Because each database is a separate file, it is easy to add it to the project, or download them from the internet later.

To convert the existing MySQL databases to SQLite, I wrote a PHP script. It queries the names of the databases on the MySQL server, then sequentially opens them. After that it makes a query for the tables in the database. The script runs a table creator query on the SQLite database, based on the lists the fields' name and type. This data is

queried by the DESC `tablename` command. After the tables are created, the data are copied by a SELECT command on the MySQL server and an INSERT INTO command on the SQLite database.

The queries on the SQLite database take long time, because each one writes the data directly to the hard drive. This process took 10-15 minutes, with constant usage of the HDD. Therefore, I installed a software called RAMDisk, which creates a virtual drive in the working memory. I changed the script to create the database on this virtual drive, and the process took only 30 seconds. After that, I copied the files to the hard drive, and added them to the project.

B. Adding the graph build and search algorithm

It was easy to add the graph build and search algorithm to the project, because it was written in Java language, which the Android Studio uses too.

On the desktop computer, I used MySQL as data source for the algorithm, with JDBC connector. The Android system handles the SQLite database natively, with its own API. To make the graph algorithm portable, I made an abstract class named Model with the necessary functions, then made a MySQLModel and an SQLiteModel class, which extend the Model class. The graph algorithm uses a variable with Model class, initialized in the main program on in the main activity with the proper constructor.

C. Graph algorithm optimization

I had to optimize the graph build and search algorithm on the desktop computer, because a usual search took more than 20 seconds.

The following optimizations were made:

- Build only the part of the graph which is between the timespan of the travel.
- Build only the part of the graph where the bus lines go through either the start or the end station.
- Store the walked path in a tree structure, rather than lists, which uses less memory and less process time.

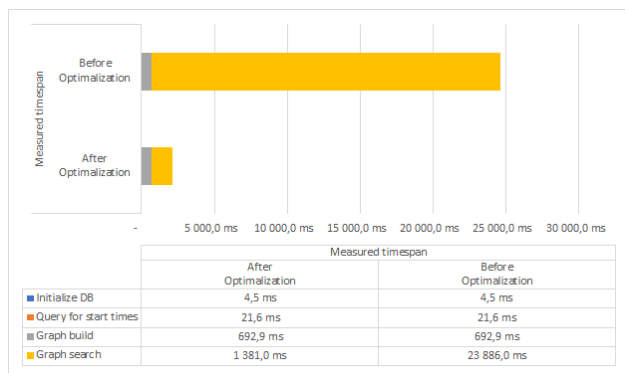


Figure 7. Algorithm runtime before and after optimization on desktop

The optimization made a good result as seen in Figure 7.

The algorithm were copied to the mobile application, and a search were triggered by a button with preset

parameters. The algorithm ran for 48 seconds, and the 90% of this were the queries within graph build.

When the graph is building, first it queries all the depart times of the buses, then queries the route of the bus for every depart time. This way there were as many queries as many depart times in the database.

I modified it to make only one query, which gets every lines' every stop in the schedule in ascending order by line id and arrive time. This way the build time became less than a second.

D. Graphical User Interface

On Android, to make the development faster and efficient, the visual interface design and the coding are separated. There is a visual editor to create the application's user interface. For the coding, the editor is an IntelliJ IDE variant. In the code, the controllers are accessible though their ID, and event handlers can be attached to them.

First, I created the main activity, added the visual controllers to the GUI, and wrote the code for it. When new activities must be started, I created them the same way. The main activity checks if there is a new database available on the web server, and download it, if it is necessary. It also opens the database connection, and initializes them for the graph build and search algorithm.

The long running tasks are called asynchronously, therefore they can run independently from the GUI, thus it can always respond to the user's input.

V. USER FEEDBACKS

The first version was uploaded to Google Play application store, and private beta test was created with 10 volunteers. The feedbacks were all positive, and they suggested new functions.

A. Minimal value for change time

Buses can come late due traffic. The minimal change time between lines were set to 1 minute by default, but the routes with such small waiting time are often impossible. I have added an input, where the user can change this setting, and adjust it to the to the first bus. (Figure 8.)

B. List of favourite stations

Passengers usually travel between a small number of stations, for example home, work, etc. The user can mark stations as favorite, and then later can select stations from this short list. (Figure 9.)

C. Shortcuts on the Main Activity

The three most used menus (route plan, stations, lines) can be accessed quicker, with a shortcut on the Main Activity. (Figure 10.)

D. User friendly display with cards.

There is no point create too many visual elements for a list or table on Android system, because they obtain a lot of resource. Instead, the RecyclerView controller reuses the containers scrolled out of the screen, reinitializes and adds them to the bottom of the list. Such container is CardView, which uses a good design to separate visual elements. (Figure 11.)

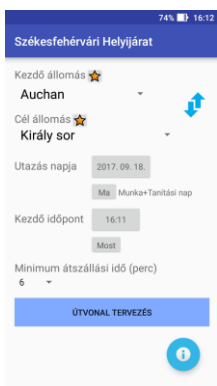


Figure 8. Minimal value for change time



Figure 11. User friendly display with cards



Figure 9. Adding stations to favorite list

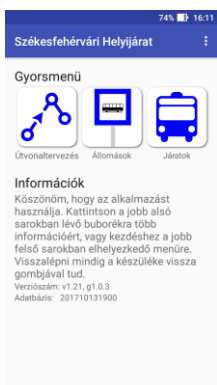


Figure 10. Shortcuts on the Main Activity

VI. CONCLUSION

I successfully designed and developed an algorithm which is capable to create a mathematical structure for a search algorithm, based on the schedule of Székesfehérvár’s local transportation. The route-planner algorithm proceeds on the structure by the given rules and creates an output, which visualizes the recommended route for the user in an easily understandable way by using various visual elements. I succeeded to create a proper input and output interface for the algorithm on mobile devices. It was glad to hear the positive feedbacks from my friends, and declared it very useful, because this information is not available on mobile devices easily, or at all.

Though the application is ready to use, I must monitor the transportation company’s website, to follow the changes in the schedule, and upload the modified databases to the webserver.

REFERENCES

- [1] StatCounter Global Stats, “Mobile Operating System Market Share Hungary | StatCounter Global Stats,” Online <http://gs.statcounter.com/os-market-share/mobile/hungary>, downloaded: 2017. 10. 05.
- [2] StatCounter Global Stats, “Mobile & Tablet Android Version Market Share Hungary | StatCounter Global Stats,” Online <http://gs.statcounter.com/os-version-market-share/android/mobile-tablet/hungary>, downloaded: 2017. 10. 05
- [3] Google Inc. *Android Developers - Dashboards*. Online <https://developer.android.com/about/dashboards/index.html>, downloaded: 2017. 10. 05
- [4] P. Schönhofen., “Nyílt forráskódú adatbázis-kezelők” (*Open source database management*) SZAK Kiadó Kft., Bicske, 2007, CD attachment: /kiegeszites/3.05.fejezet.pdf.
- [5] S.K. Aditya, V.K. Karn, *Android SQLite Essentials*, Packt Publishing, 2014

Applying Structure-from-Motion technique for visual odometry

V. Potó and Á. Barsi

Budapest University of Technology and Economics, Budapest, Hungary

potovivien@epito.bme.hu, barsi.arpad@epito.bme.hu

Abstract — Autonomous vehicles have several sensors, that enable sensing their environment. Positioning with GNSS systems has limits regarding the availability/accessibility and accuracy. The positioning accuracy of these vehicles can be increased using the environmental sensors. Karlsruhe Institute of Technology (KIT) and Toyota Technological Institute at Chicago have created a test data set, named KITTI containing color and grayscale camera image series, lidar measurements and GPS/INS data. We study the suitability of these image data for positioning purposes. In our paper we calculate the positioning by visual odometry.

I. INTRODUCTION

As well as the world is developing so fast, people's needs grow also. There was a huge development in the last 200 years in the evolution of cars. The first car was designed by François Isaac de Rivaz in 1808. It worked with internal combustion engine and with hydrogen. In 1870 the first four-cycle, gasoline powered combustion engine came out, that was made by Siegfried Marcus. Nikolaus Otto invented the four-stroke petrol internal combustion engine and Rudolf Diesel made the four-stroke diesel engine. The beginning of battery electric car was bounded to Ányos Jedlik and Gaston Planté. [1]

In the last years a new era in the evolution of cars can be distinguished: the development of autonomous vehicles. According to Wikipedia, "An autonomous car is a vehicle that is capable of sensing its environment and navigating without human input." [2]

This new milestone brings new problems and questions that the world should solve. One of them is the localization and navigation. Nowadays the drivers use GNSS systems for positioning, but for example in urban environment with the tall buildings the satellites are invisible and measuring less than four satellites, the positioning fails. Furthermore, even if the number of satellites is enough, the car has problems with the accuracy. There are two choices: GNSS systems must be improved to have higher accuracy, or other additional methods have to be involved. One of them is the use of highly accurate and detailed map, which contains the road and its surroundings. Vehicles of nowadays are equipped with different sensors (cameras, lidars and radar-based sensors) to capture their environment constantly. Comparing the sensed data with the content of the map, the self-driving car can specify its location and plan its further path.

There are numerous methods to solve the localization problem. We have applied visual odometry computed on the bases of Structure-from-Motion (SfM) technique, so the position has been derived from the captured images of the onboard cameras. For this experiment we have used the test dataset from the KITTI Vision Benchmark Suite.

II. DATA

KITTI is an exciting project in Karlsruhe, Germany, created and managed by the Karlsruhe Institute of Technology (KIT) and Toyota Technological Institute at Chicago. A Volkswagen Passat B6 was equipped with the following instruments:

- 1 Inertial Navigation System (GPS/IMU): OXTS RT 3003,
- 1 Laserscanner: Velodyne HDL-64E,
- 2 Grayscale cameras, 1.4 Megapixels: Point Grey Flea 2 (FL2-14S3M-C),
- 2 Color cameras, 1.4 Megapixels: Point Grey Flea 2 (FL2-14S3C-C),
- 4 Varifocal lenses, 4-8 mm: Edmund Optics NT59-917.

The system configuration is in Figure 1 and Figure 2.

There were dozens of urban, rural ways as well as highways captured by the probe vehicle in Karlsruhe and the surroundings. An eight core i7 computer with a RAID system, running Ubuntu Linux and real-time database was used to record the obtained data. All cameras were directed forward.

The freely available dataset can be used for several purposes: stereo image processing, optical flow, visual odometry, 3D object detection and 3D tracking. The researchers are supported by evaluation metric, so new, improved methods can be validated. [3]

We have used the raw data and the odometry dataset. The raw data has been classified into the following categories: city, residential, road, campus, person and calibration. The zipped datasets contain four data packet: unsynced + unrectified data, synced + rectified data, calibration files and tracklets. The difference between the unsynced + unrectified data and synced + rectified data is the level of processing. In the second data group all images are already undistorted and rectified, and synchronized with all sensor observations via time stamps. The first dataset contains the raw data, obtained directly from the instruments. Tracklets are elementary labelled objects (e.g. car, truck, tram, pedestrian) along a track.

We have chosen the synchronized and rectified dataset for our research work considering different environmental types. These datasets are presented with some features in Table 1.



Figure 1. Fully equipped probe vehicle [3]

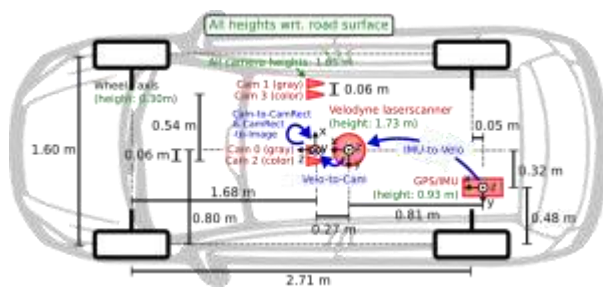


Figure 2. Fully equipped probe vehicle – top view [3]

Each dataset is labelled with a short dataset number. The full dataset identifier consists of the date of measuring, the word “drive” and the sequence number, e.g. 2011_09_26_drive_0020, which is the 20th dataset recorded on 26th September 2011. All zipped dataset stores the following file structure in 6 directories: image_00, image_01, image_02, image_03, oxts and velodyne_points. The first four folders are for the 1392 × 512 pixel sized camera images in png format and a text file with timestamps; the first two directories contains the grayscale images, then the color images, respectively. The oxts folder contains the GPS and IMU data in txt format and a description of the data format and timestamps. The Velodyne_points folder contains the lidar data points in binary format and 3 timestamp files. We have used solely the color images of the folder “image_02” and the oxts directory for validation.

A development kit with Matlab and C++ codes is also available to help people in the use of the data sets. [4, 5]

III. METHOD

The term odometry has the meaning of „route measurement” coming from composing two Greek words. In robotics, it is meant as the estimation technique of positioning of a wheeled robot relative to a starting location. It has been realized mostly by the measurement of wheel rotation (The so-called rotary encoders obtain the information e.g. in cars about the distance travelled). If the diameter or radius of the wheel is known, the distance can be calculated by multiplying the perimeter by the number of rotations. Beside this technique, odometry is continuously studied and visual odometry has also been invented. Visual odometry is per definition „the process of

 TABLE I.
MAIN PARAMETERS OF THE EVALUATED DATASETS

Dataset number	Short dataset number	Category	Number of photos	Shape
2011_09_26_drive_0020	20	residential	420	arched
2011_09_26_drive_0032	32	road	390	hooked
2011_09_26_drive_0039	39	residential	395	straight
2011_09_26_drive_0070	70	road	1104	arched
2011_09_26_drive_0093	93	city	433	broken
2011_09_26_drive_0095	95	city	268	arched
2011_09_26_drive_0104	104	city	312	straight
2011_09_26_drive_0117	117	city	660	hooked
2011_09_28_drive_0001	1	city	106	arched
2011_09_29_drive_0004	4	road	339	straight
Odometry dataset	Odo	city	531	broken

determining the position and orientation of a robot by analyzing the associated camera images” [6]

Visual odometry requires therefore camera images being suitable for computing the position of the vehicle. Because cameras are important components of the future’s vehicle, big efforts have been taken to fix cameras on vehicles, to capture images and to evaluate them in order to support the vehicle control, mostly to detect obstacles, pedestrians and other vehicles on the road. The basic idea with visual odometry was the usage of these captured images also for deriving the position. Among the wide spectrum of possible solutions, our approach was focused on a mature technique applied in digital photogrammetry and image analysis [7].

The Structure-from-Motion technology is based on point features, which can be detected on the images and have exact identifiable position. There are interest operators, e.g. Förstner, Harris, Moravec, SIFT, SURF etc., which are standard tools in computer vision and extract points in images being in corners, intensity jumps, line ends. If the so determined points can be found in multiple images, these points give the possibility to couple the images in space (Figure 3). The more points are exactly identified and found in several images, the better merge can be achieved. The merging step is done pairwise and at the end a bundle adjustment can “fine tune” the whole system. After aligning all obtained images, the result is the relative orientation elements of the image projection centers. It means that the first image defines the coordinate system, in which the second image is coupled to the first one, then the third to the second one and so on. The coordinates of the image projection centers are showing exactly the movement of the camera carrier platform, i.e. the trajectory of the vehicle in a local reference system.

The described image alignment is an elementary part of an object reconstruction software package, like Agisoft Photoscan, Pix4D, VisualSFM. The basic idea of our paper was to test the ability of such software in the solution of the visual odometry problem related of vehicular camera images.



Figure 3. A Points obtained by interest operator being at least on 13 images

IV. RESULTS

The positioning calculation of visual odometry has been conducted by Pix4D Mapper run in the cloud. The Amazon hosted virtual machine had two Intel(R) Xeon(R) CPU E5-2666 v3 @ 2.90GHz processors, 36 threads, 60 GB available RAM and Linux 3.13.0-91-generic x86_64 operating system (There was no need for GPU power).

The resulting image projection centers in green can be seen with some spatially determined tie points in Figure 4. All of the projection centers visualized forms a sequence, where the probe vehicle has moved. This sequence is therefore corresponding to the trajectory of the vehicle. In the test set the vehicle’s navigation data is also available: the GPS and inertial measurements were fused and are also available to analyze the vehicle’s movement. In our project this GPS-based data was applied for validating and

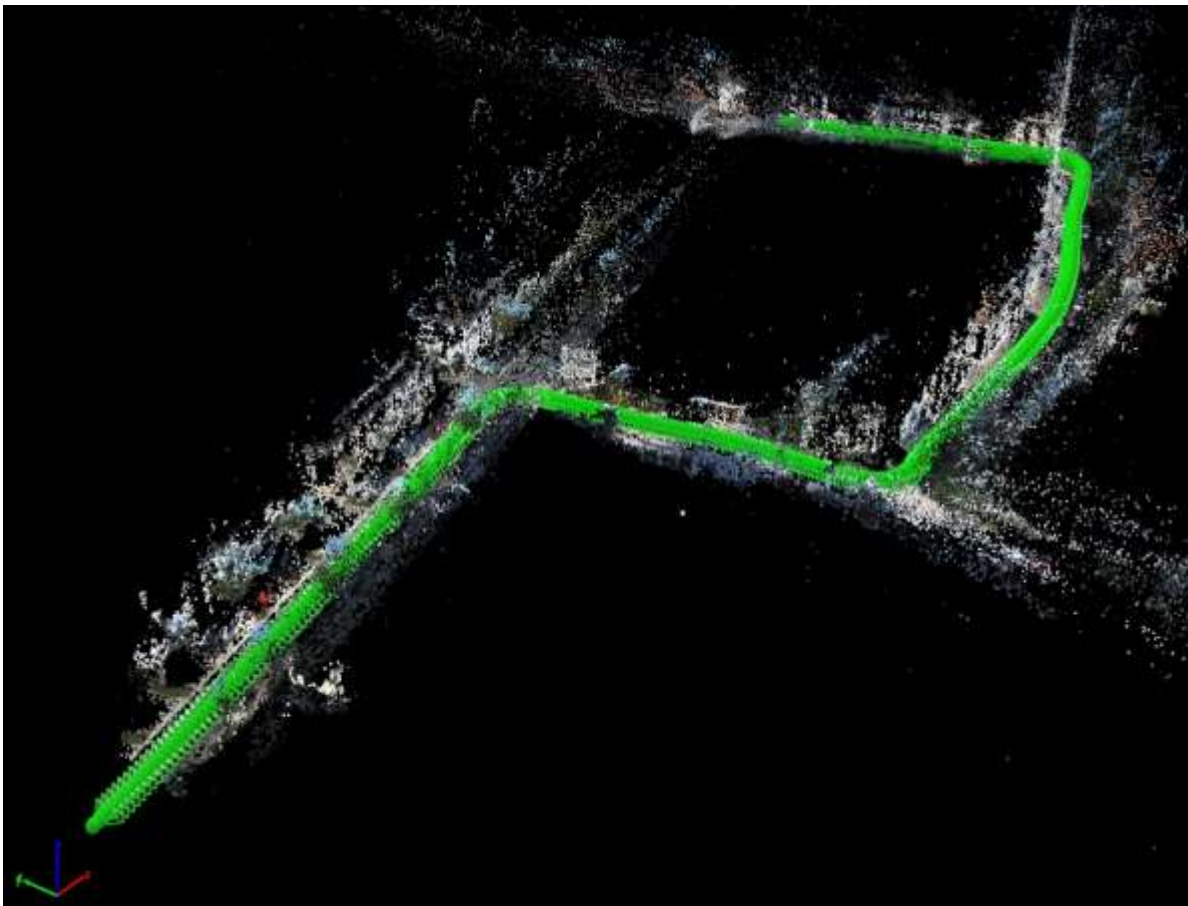


Figure 4. Image projection centers and image frames of Odo dataset in Pix4D environment

checking the quality of the obtained odometry measures. Mathworks Matlab was used for all further analysis steps.

Odometry produces the projection center coordinates in a local coordinate system. The center points have to be transformed into WGS84 geographic coordinate system, where the GPS/IMU positions (stored in oxts file) are given. Using common points of both coordinate systems the necessary transformation parameters can be achieved.

In Figure 5 we can see the oxts dataset (reference data) in green, and the transformed odometry data in blue. Figure axes represents the longitude and latitude coordinates in decimal degrees.

Some differences between the two data sequences can be noticed. There are breakings in the sequences, where the continuation of the sequences could not be achieved in odometry, i.e. there were some problems with merging the corresponding images. Of course, similar gap can be detected also in the reference measurements, meaning that the GPS signal receive had also troubles.

One can calculate an estimation for transforming the geographic coordinates into planar metric system. Then the two sequences can be compared; not only visually, but numerically, too. The point-by-point computation of differences between the odometry and reference data can be seen in Table II. The first column of the table shows the dataset short identifiers as numbers, in the next columns are some statistic data for the differences: the average

Table II.

Statistics of results of the evaluated datasets in meter

Short number	Average difference	Maximal difference	Difference's median
1	2.6591	4.9502	2.7608
4	18.9735	24.782	20.4966
20	-	-	-
32	58.8058	93.2657	64.207
39	4.9704	17.4699	4.3955
70	2.6282	3.4853	3.1438
93	69.2613	196.2815	43.505
95	21.1573	49.1437	17.5812
104	1.4138	2.7006	1.3488
117	3.2015	5.8808	3.5387
Odo	0.6710	1.0749	0.6611

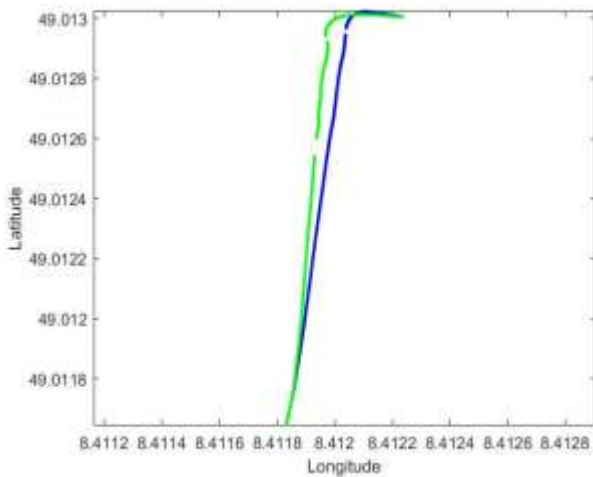


Figure 8. The calculated trajectory (Dataset 117). (Blue – result of the visual odometry, green – GPS/IMU reference)

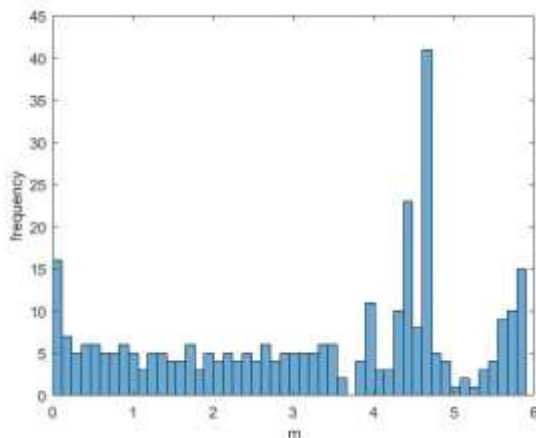


Figure 9. Histogram of differences for Dataset 117



Figure 5. The result of Dataset 4 with black line in Matlab Webmap environment



Figure 6. Dataset 93 in Matlab Webmap environment



Figure 7. Dataset 39 in Google Earth environment

difference, the maximal difference and the difference's median. All given features are in meter. There was some unacceptably high difference, they are marked in red and are removed from the further analyses. [8]

For Dataset 20 there is no statistics, because there was a mistake in the dataset, so it has been dropped.

There are some extremities in the differences in the case of Dataset 93. With this dataset the GPS measurements were fully chaotic, some strange errors occurred. The odometry has resulted a smooth, believable trajectory, that can be seen in Figure 8, but still this dataset was also dropped.

With Dataset 32 – straight trajectory in urban environment – the obtained differences are mostly acceptable, but approaching the midpoint of the track, this difference reach almost 100 m. After the midpoint the computed trajectory goes back to the normal state (arched form differences). Supposedly in this case there were some point identification problem and systematic error in merging.

Considering Dataset 95 there is a similar effect in image coordinates. This dataset's shape is also arched. Figure 6 presents a histogram of differences for Dataset 117. On the x axis there are the differences in meter, on the y axis, the frequencies, respectively. The distribution of the differences is almost equal, only two spikes are visible between 4 and 5 m. This figure illustrates that the SfM based visual odometry can result stable solutions.

We have used also another methods for displaying and verifying. The transformed odometry datasets can be shown with OpenStreetMap background, so the interpretation of the trajectories is easier. An example can be found on Figure 7 for Dataset 4 with black line on the beltway. Some noise can be detected at the upper end of the black line.

Another presentation way is with Google Earth environment. In Figure 9 there is the Dataset 39 in green with the verifying oxts data in red.

V. SUMMARY

The 3-dimensional object reconstruction has been solved by the Structure-from-Motion technique. This algorithm has a processing step, when the projection centers of the taken images are calculated. The result is a coordinate tuple in a local reference system.

Because the development of the future autonomous vehicles is strongly based on camera images, numerous equipment is developed to collect imagery. The combination of the image processing of the automotive (onboard) cameras and the photogrammetric object reconstruction by the Structure-from-Motion technique

was the basic idea of our research. The results have shown that the projection center calculation is a possible way for the visual odometry solution.

The test can point on that the captured images of photogrammetrically lower geometric resolution are suitable to execute this task. The image set collected during a test drive has many images, but the content change is quite low between consecutive images, so the similarity enables their coupling and the relative position could be derived.

In the test suite several environmental types (residential, city and road) were chosen. The applied technique can be evaluated in this aspect, too. The prior expectation was that city has the highest variability, has the highest amount of image patterns, the most features can be detected there and so the best solutions will be there. Our most interesting trajectory is also in a city (Odo). This hypothesis was not checked directly but the city test data were mostly successfully evaluated. Of course, if city or residential roads are used and the cameras captures trees or similar objects along the trip, enough patterns and features are present to be able to solve the odometry challenge.

Furthermore, if the odometry is solved, the spatial positions of the cameras are known, then using two or more corresponding synchronized cameras opens the way to apply stereo object positioning and reconstruction. This achievement is essential in environmental sensing and the vehicle control, when other vehicles, pedestrians or any objects on the road (like dropped cargo) must be detected and accidents can be avoided.

REFERENCES

- [1] Wikipedia – History of the automobile
https://en.wikipedia.org/wiki/History_of_the_automobile
- [2] Wikipedia – Autonomous car
https://en.wikipedia.org/wiki/Autonomous_car
- [3] KITTI webpage
<http://www.cvlibs.net/datasets/kitti/index.php>
- [4] Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (n.d.). "Vision meets Robotics: The KITTI Dataset"
<http://www.cvlibs.net/publications/Geiger2013IJRR.pdf>
- [5] Geiger, A., Lenz, P., & Urtasun, R. (n.d.). "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite."
<http://www.cvlibs.net/publications/Geiger2012CVPR.pdf>
- [6] Wikipedia – Visual odometry
https://en.wikipedia.org/wiki/Visual_odometry
- [7] Somogyi Á - Barsi Á Pixel-based 3D Object Reconstruction, In: Orosz Gábor Tamás (Ed.) 11th International Symposium on Applied Informatics and Related Areas (AIS 2016), Székesfehérvár, 2016.11.17, pp. 60-63
- [8] G. Mélykúti, "Topography – Basic terms in mapping" in Hungarian 2010
http://www.tankonyvtar.hu/hu/tartalom/tamop425/0027_TOP1/ch01s04.html

Use of Drones for Data Gathering in Precision Farming

M. Veróné Wojtaszek* A. Tangl**

* Institute of Geoinformatics, Alba Regia Faculty of Engineering, Óbuda University, Székesfehérvár, Hungary
wojtaszek.malgorzata@amk.uni-obuda.hu

** Institute of Business Sciences, Faculty of Economics and Social Sciences, Szent István University, Gödöllő, Hungary

Abstract— Precision agriculture is a farming management concept based on observing, measuring and responding to inter- and intra-field variability in crops. The goal of precision agriculture is to more efficiently apply a farm's limited resources to gain maximum yield. For doing that we need precise and frequent data about the condition of crops. Drones offer a very efficient way of gather data over a large scale operation. UAV offers farmers the chance to get easy access to relevant information out of a large amount of high quality imageries. In this study the professional mapping drone (eBee) was used to get data about vegetation. The papers give an overview of available multispectral cameras used to get data of plants from drone platform. We examined the potential of utilizing an UAV for the characterization and monitoring of cultivated land. The vegetation indices have been used to information extraction from imagery.

I. INTRODUCTION

Modern agriculture must rely on expert knowledge and implement new technology to enable farmers for profitable production while fulfilling environmental and food safety conditions. The traditional methods of cultivation have not taken into consideration the variability of habitat conditions. The doses of fertilizers and pesticides are determined for the mean conditions of the field. During the traditional cultivation a large variation in mineralization in a given field resulted local overdose on rich soil or poorly fertilized on poor soil.

Precision agriculture technology is a farm management system, which relies on various measurements, data collections and analysis, as well as decision making. Measurements include soil chemical and physical characteristics determination, grain yield and quality measurements, and several remotely sensed property determination as well.

Precision agriculture (or precision farming) is a collection of agricultural practices that focus on specific areas of the field at a particular moment in time. This is opposed to more traditional practices where the various crop treatments, such as irrigation, application of fertilizers, pesticides and herbicides were evenly applied to the entire field, ignoring any variability within the field.

Research concerning precision agriculture in Hungary started in the late 1990s [1]. In this initial phase, the study focuses on economical questions of the technology, such as return of investment or efficiency of the technology. In the other research authors presented technological development possibilities of precision agriculture. During their work economic circumstances (profitability) were

mapped in a special profit-map [2]. Among other influencing factors on decision making. Parallel to these calculations optimal amount of fertilizer (nitrogen) in the given economic circumstances were calculated, where profit and energy balance would have been the best (optimum). Numbers of the published works are investigating precision plant protection issues. It is important to mention that in the period prior to the 1990s several researches had examined the soil fertility potential and relationship between fertilization and yield as well. Berzsenyi and Györfly [3] have shown that yields depend on two factors: nutrition (30.7%) and genetic soil type (30%). However, crop yields can be greatly affected by soil degradation. In the case of site-specific cultivation, the nature of degradation within agricultural area should be take into consideration [4]. A number of processes of degradation threaten soil functions and reduce fertility of the area. With the expansion of technology, several studies have been published that approach the precision management in a complex way, 'Precision Crop Production' [5], 'The Methodology of Precision Agriculture' [6].

The introduction of precision technologies in agriculture has been motivated by the high degree of variability of agro-ecological conditions within fields. One of the criterion for introducing precision agricultural technologies is the development of an up-to-date arable crop information system that provides information on soils, crop land cultivation, plant status, etc. This information can be used as starting data for cultivation, for predicting yield estimate. In order to set up such an information system, it is essential to use modern data gathering and analysis technologies. Remote sensing is the most effective tool for surveying the Earth's surface and tracking its changes.

The precision agriculture is a possible way to optimize the economic resources. The accounting system reflects the changing of revenues, costs and inventory. The different productivity improvement methods decrease the waste and can optimize the level of inventory [7].

II. DATA GATHERING BY UAS

Recently, a new data gathering technology so called UAS has been released. Data collection via unmanned aerial systems (UAV) is a beneficial service for the agricultural sector. The use of these lightweight aircraft, operated by a remote pilot or autonomously through programming, provides a faster, safer and more economical means of collecting valuable data. Growing use of UAV imagery offers farmers and agronomists the

chance to get easy access to individual relevant information out of a large amount of high quality imagery. These systems, commonly known as drones, can be equipped with hyperspectral or RGB cameras to capture many images of a field that can be processed to create orthophotos and NDVI maps.

We have used the professional mapping drone (eBee) to capture high-resolution aerial photos. The images can be transformed into accurate orthomosaics & 3D models. During the image analysis the vegetation indexes were created. In our study the NIR camera was applied to capture specific imaging bands in the near-infrared range. This data was then run through a special processing algorithm to create the NDVI (Normalised Difference Vegetation Index) imagery, being the standard for documenting and assessing crop and vegetation conditions and health.

A lightweight UAV platform (eBee), which is developed by senseFly was used throughout this study. The eBee is a fixed-wing UAV that weighs less than 0.70 kg, including the camera, and has a wingspan of 96 cm. Its cruising speed ranges from 40 to 90 km/h, which makes it suitable for mapping up to 12 km² (1200 ha) with a maximum flight time of 50 minutes.

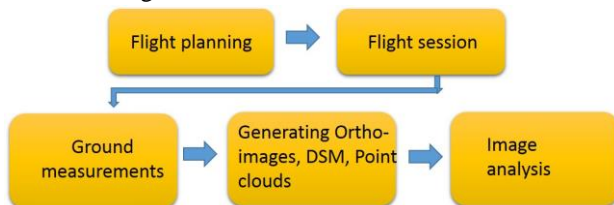


Figure 1. Study workflow

In the 'Flight Planning' stage, preliminary parameters of UAV flight, were evaluated based on area covered by the test area, wind velocity, atmospheric conditions, etc. During the 'Flight Session' and the 'Ground Measurements' stage the flights took place and ground control point (GCP) measurements were work were done. In the stage 'Generating Ortho-Images, DSM and Point Clouds,' a semi-automated process was conducted.

The main objective of this study was to evaluate the effectiveness of UAV in mapping vegetation conditions. OBIA and NDVI based classification approach was used to map the crop.

III. INDICATORS OF VEGETATION AND ECONOMIC BENEFITS

By measuring the reflectance of the plants at various wavelengths, it is possible to collect a lot of information about the status of the plants. From these data we can derive vegetation indices that help us estimate the vegetation cover and the Leaf Area Index (LAI). As the population of plants increases, the amount of biomass causes an increase in the overall near-infrared reflectance and a reduction in the red reflectance. From previous research, there are known relationships between the indices using those two regions of the spectrum and the amount of vegetation, a measure of which is the Leaf Area Index. From these estimates we can derive the population of the he data provided by the sensors can also be used to make an estimate on the future crop yield. By calculating the Normalised Difference Vegetation Index (NDVI), or

the Soil-Adjusted Vegetation Index (SAVI) when vegetation cover is low, we can get information on the crops' vigour. Low index values usually indicate little healthy vegetation while high values indicate much healthy vegetation. Different indices have been developed to better model the actual amount of vegetation on the ground. A lot of research has been done to derive relationships between a vegetation index of a crop, measured at a particular time, and the final crop yield.

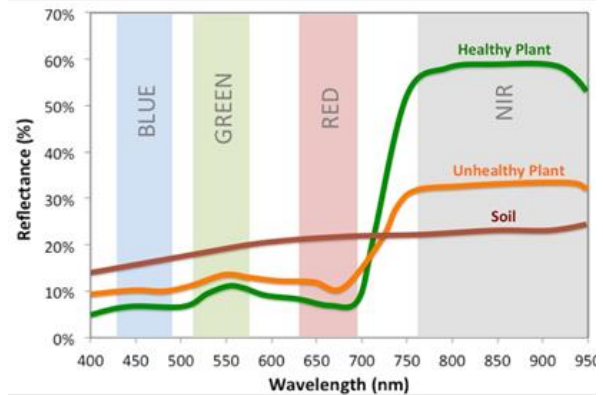


Figure 2. Spectral curve of the light reflected from the plant (<http://static.micasense.com/img/reflectance2.jpg>)

IV. CLASSIFICATION

Crop monitoring is the one of these applications since remote sensing provides us accurate, up-to-date and cost-effective information about the crop heterogeneity at the different temporal and spatial resolution. In this study classification was performed with different kind of vegetation indices. The rule set (algorithm) was developed to identify crop heterogeneity within field using the 6 cm spatial resolution UAV images. The process considered spectral value and spatial characteristics of objects and it is included the following steps: segmentation, feature extraction and object classification. In this process, the images were segmented into homogeneous multi-pixel objects using the multiresolution algorithm. Segmentation is a bottom-up region-merging process in which the image is subdivided into homogeneous objects on the basis of several parameters (band weights, scale, color, shape, smoothness and compactness) defined by the operator. Two levels of segmentation were used throughout the procedure: identification of study area (chess board segmentation) and the second level to generate smaller objects for mapping crop conditions. The classification results can be used to determined management zones to get an optimal amount for each input in crop production, founded on variability of soil characteristics and the other factors conditioning a crop yield (Fig. 3.). In this case, we define locations within field with the same or very similar conditions for crop planting. The figure below (Fig. 3) shows the effect of erosion on plant growing (winter wheat field in autumn and spring). As wheat evolves, it is increasingly marked difference between the eroded area and the area not affected by erosion.

The information about the vegetation will influence the use of economic resources. To optimize the fertilizer, the pesticides, the cultivation etc. is the financial advantage of quick automated drone imaging. Computing the right

amount of resources reduce the expenditures. The differential use of supplies balances the yield. Due to the drone pictures the agricultural waste decreases, the revenues thru the additional product increase and the level of inventory can be minimized. The operation and maintenance cost of the machines can be reduced because they are used only on the necessary areas what are indicated by the images.

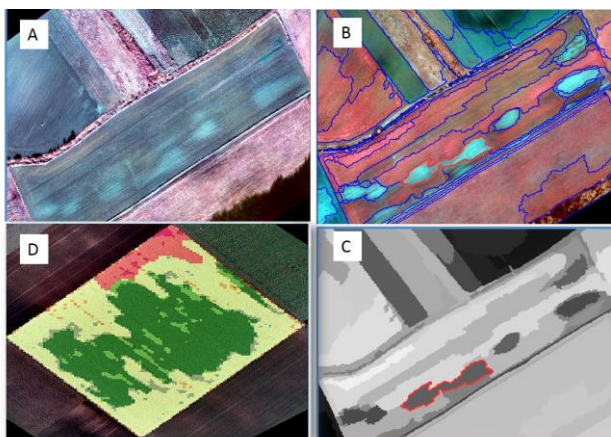


Figure 3. Effect of erosion on plant growing: winter wheat field in autumn (A) and spring (B). C:NDVI image, D: management zones

V. CONCLUSION

Drones can help farmers catch problems faster and react more quickly, which can save money in crop losses per field and the data generated by drones help farmers gain a more accurate and detailed picture of how their crops are reacting to their management strategies, which can lead to more effective use of resources. By using the precise drone data the farmers can calculate the company efficiency based on the accounting information. This contributes to the better planning of company revenues and expenditures, reducing stocks, and increasing the profitability and free cash-flow. The reduced resources and increased productivity can optimize the use of the

land surface. The quicker information leads to quicker and better financial decision possibilities and preventing wasteful actions.

Spectral indices derived from satellite data are widely used for land cover change research. They can reduce the data volume for analysis and provide combined information that is more strongly related to changes than any single band.

Vegetation indices use various combinations of multi-spectral satellite data to produce single images representing the amount of vegetation present, or vegetation vigor. Low index values usually indicate little healthy vegetation while high values indicate much healthy vegetation. Different indices have been developed to better model the actual amount of vegetation on the ground.

REFERENCES

- [1] B. Györfly.: 2000. Javaslat a precíziós agrárgazdálkodás kutatási programjának indítására. A Magyar Tudományos Akadémia Agrártudományi osztályának 2000. évi tájékoztatója. Budapest, pp. 17-22.
- [2] N. Smuk et al: 2010. Jövedelemtérképek a precíziós növénytermelésben. *Gazdálkodás* 54. évfolyam, 2. szám. pp. 176-181.
- [3] Z. Berzsényi – B. Györfly.: 1995. Különböző növénytermesztési tényezők hatása a kukorica termésére és termésstabilitására. *Növénytermelés*, 44. 507-517.
- [4] M. Verőné Wojtaszek: 2011. Földhasználati tervezés és monitoring. TÁMOP jegyzet. 4.1.2-08/1/A-2009-0027
- [5] T. Németh: Precíziós növénytermesztés (részjelentés 2002 okt.), MTA Talajtani és Agrokémiai Kutatóintézet
- [6] T. Németh et al: A precíziós mezőgazdaság módszertana, JATEPress-MTA TAKI, Szeged 2007
- [7] A. Vajna Istvanne Tangl – I. Vajna: 2014. Relationship between Inventory and Value Stream Management, In *New Trends in Management in the 21st Century*, Czestochowa: Czestochowa University of Technology, 139-154.

Camera Calibration and Semi Global Image Matching

Tamas Jancso

*Institute of Geoinformatics, Alba Regia Technical Faculty, Óbuda University
Pirosalma u. 1-3, Székesfehérvár, H-8000, Hungary*

jancso.tamas@amk.uni-obuda.hu

Abstract— the paper deals with the building of 3D models with photogrammetric image matching based on stereo-pairs. The semi global matching (SGM) is widely used in photogrammetry for building 3D models from stereo-pairs. The quality of the produced model depends on many factors including the SGM modes, minimum-disparity, maximum-disparity, block-size, and uniqueness-ratio. The experiments showed us that the quality and accuracy of the camera calibration directly influences the quality of the output disparity map. The other important issue is the ratio between the photo-base and the distance of object points. Increasing the photo-base the disparity map quality is decreasing. The example calculations are carried out in MATLAB.

Keywords— SGM, camera calibration, image matching, disparity map

I. INTRODUCTION

For testing, a stereo-pair was used to build a disparity map, which enables to build the 3D model later. A Sony Alfa A350 digital SRL camera was used, which was calibrated by chessboard test field. The calibration procedure was carried out in MATLAB application [3]. Using the camera calibration data stereo-pair of the taken scene was oriented based on common points. The common points were identified automatically using the SURF function in conjunction with the appropriate gross-error filtering methods. The translation vector and the rotation matrix were calculated as the elements of the relative orientation. Based on these elements the essential and the fundamental matrix were generated as well. After this the images were transformed into a normal case image to have parallaxes only horizontally. This rectification procedure enables and makes easier the identification of common points for generation of the disparity map using the SGM (Semi Global Matching) image matching method. The result of the disparity map depends on the setup of the initial parameters. The most sensitive parameter is the disparity range. The disparity range can be adjusted only with manual measurements which makes the SGM method not a full automatic process. The whole procedure is summarized on Fig. 1.

II. TEST SCENE

For testing, an artificial scene was used with some objects having rectangular and spherical forms in different colours and textures as it is seen on Fig 2. The background was relatively flat

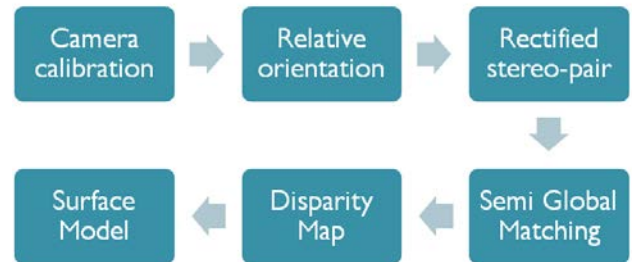


Fig. 1 Procedure of the image matching process



Fig. 2 Test scene

The object distance was about 1 m with a basis of 20 centimetres. Which means the ration between the basis and the object distance was 0,2. Two photos were taken close to normal case.

III. 3. CAMERA CALIBRATION

For the further calculations and orientation procedure we need the intrinsic date of the applied camera. The calibration procedure was carried out in MATLAB using the built-in Single Camera Calibration application. During the calibration the following parameters were calculated:

- Radial distortion parameters (K1, K2 and K3)
- Tangential distortion parameters (P1 and P2)
- Skew parameter (s)

These parameters describe the distortion effect of the lens system and the skew of the sensor area.

A. Preparation

The photos were taken with a Sony Alfa A350 DSRL camera. For camera calibration, a chessboard test-field was used as it is seen on Fig. 3.

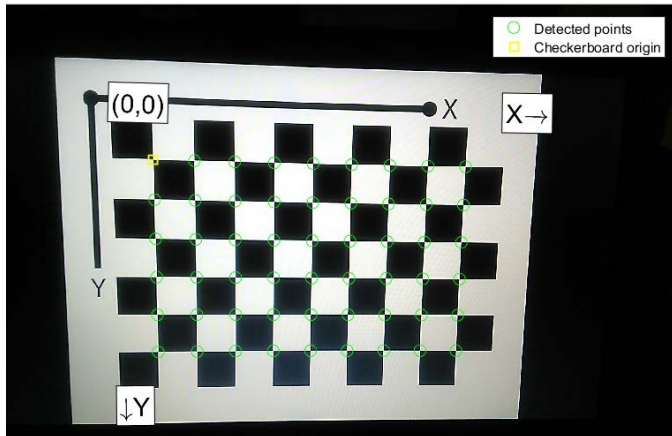


Fig. 3 Calibration field

The photos were taken from different direction trying to cover the whole sensor area. During the shooting the camera body was also rotated with 90 degrees. The series of photos was filtered, and the unfocused images were dropped out.

B. Calculation of Intrinsic Parameters

After uploading the remaining images, the calibration software identified automatically the corners of the chessboard table and the intrinsic parameters were calculated with an adjustment procedure. The mean errors calculated from the residuals were displayed and the user could adjust the level of outliers manually. The Fig. 4. indicates the reprojection errors. From the figure we see that the overall mean error is around 0.22 pixels for all images.

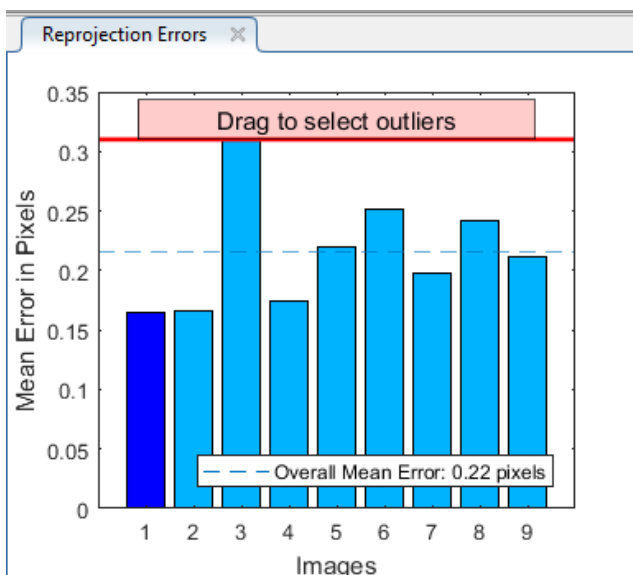


Fig. 4 Reprojection errors

The user can decide which parameters should be included in the adjustment procedure. We can try different combinations. We can choose only K1, K2 parameters or we can add K3 parameter the unknowns. Or we can include P1, P2 tangential distortion parameters, and all this can be combined with inclusion of the skew parameter. The result can be visualized in a 3D scene of the camera and the taken images (Fig. 5).

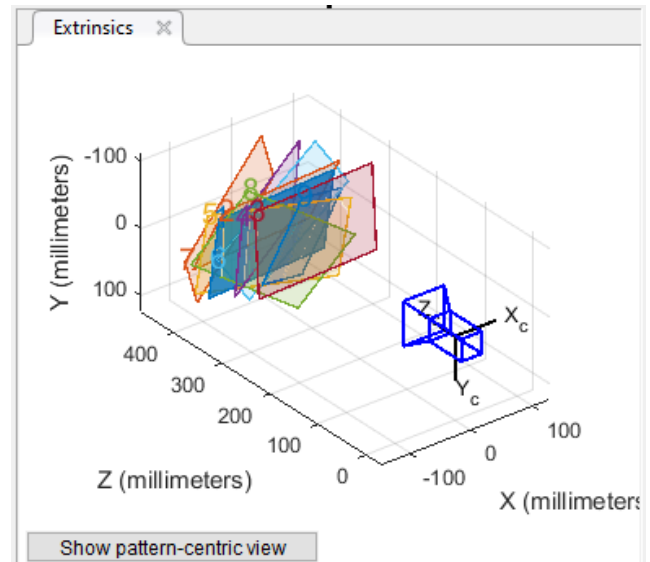


Fig. 5 Orientation of images relative to the camera position

IV. 4. RELATIVE ORIENTATION

C. Identification of Common Points

The identification and measurement of common points is a multi-step procedure on a stereo-pair of images. First, we must identify the objects itself on both photos with the help of a segment based algorithm. Here the SURF function was used to separate and identify the object points. After this the program tries to make pairs of points as common points on the stereo-pair. In this phase some matching of points can have real big errors as it is seen on Fig 6.

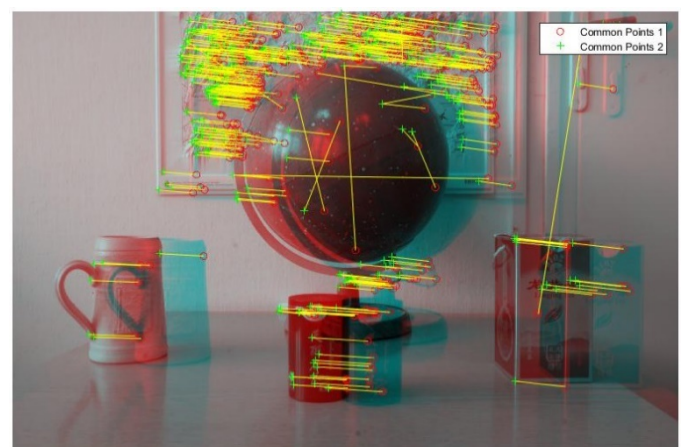


Fig. 5 Common points including matching points with large errors

The next step is the calculation of the Fundamental Matrix, which is stored as a 3-by-3 matrix. The Fundamental Matrix relates the two stereo cameras, such that the following equation must be true [3]:

$$[P_2 \ 1] * FundamentalMatrix * [P_1 \ 1]^T = 0$$

Where P_1 , the point in image 1 in pixels, corresponds to the point, P_2 , in image 2.

With a similar calculation the Essential Matrix is calculated, which is stored as a 3-by-3 matrix. The Essential Matrix relates the two stereo cameras, such that the following equation must be true [3]:

$$[P_2 \ 1] * EssentialMatrix * [P_1 \ 1]^T = 0$$

Where P_1 , the point in image 1, corresponds to P_2 , the point in image 2. Both points are expressed in normalized image coordinates, where the origin is at the camera's optical centre. The x and y pixel coordinates are normalized by the focal length of f_x and f_y .

During the calculation we try to eliminate the errors with the Least of Median of Squares method. This method uses the input points that are already putatively matched. Using the algorithm, we can eliminate any outliers which may still be contained within putatively matched points. The result is shown on Fig 6.

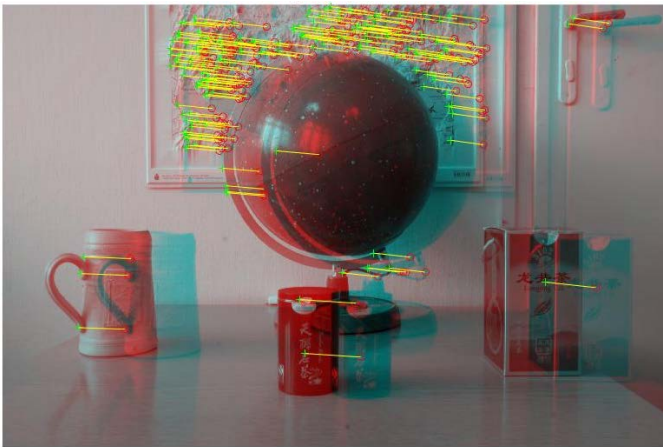


Fig. 6 Common points after eliminating the large errors

After this step we can build up the Stereo Parameters objects in MATLAB. For this we need to follow the following code:

```
1. [relativeOrientation, relativeLocation] =
   relativeCameraPose(EssentialMatrix, cameraPa
   rams_Sony, cameraParams_Sony, inlierPoints1, i
   nlierPoints2);
2. [rotationOfCamera2, translationOfCamera2] = ca
   meraPoseToExtrinsics(relativeOrientation, re
   lativeLocation);
3. stereoParams =
   stereoParameters(cameraParams_Sony,
   cameraParams_Sony, rotationOfCamera2, transla
   tionOfCamera2);
```

In line 1 we calculate the relative orientation using the previously calculated `EssentialMatrix`, the intrinsic camera parameters (`cameraParams_Sony`) and the common points (`inlierPoints1, inlierPoints2`).

In line 2 we calculate the rotation matrix and the translation vector of camera 2 relatively to camera 1.

In line 3 and 4 we finally build the `stereoParams` object which is necessary for the rectification of images into a normal case.

D. Rectification of Images

The rectification of images is necessary to build a normal stereo-pair where the parallaxes between the common points exist only horizontally. During the rectification, applying the `rectifyStereoImages` function, the distortion effects are also considered and eliminated. The result is shown on Fig. 7.



Fig. 7 Anaglyph image of rectified images

Using the Image Tool application in MATLAB we can measure the distances (horizontal parallaxes) between some common points (Fig. 8). Visually we can select and decide the maximal parallax which will be useful later when we try to setup the disparity range during the SGM processing.



Fig. 8 Measuring parallaxes in Image Tool

E. Building the Disparity Map and the 3D Point Cloud

Applying the SGM algorithm we try to match points between I_1 as left image and I_2 right image with maximum accuracy, then we are looking for the minimum value of the following energy function [1], [2]:

$$E(D) = \sum_p \left(C(p, d_p) + \sum_{q \in N_p} P[|d_p - d_q| \geq 1] \right)$$

The first member of the function summarizes the C cost spent for matching a given pixel p , where d_p is the parallax between images I_1 and I_2 . Since the matching is done by individual pixels, only the $I_1(p)$ and $I_2(q)$ intensities are considered for calculation of the cost. The second member of the function means a penalty for those q pixels where the difference in parallaxes is larger than 1 for the p pixel considering the N_p neighbourhood. It means that the P penalty member can cause an unsuccessful matching even in those case where the difference in intensity between $I_1(p)$ and $I_2(q)$ equals zero, but the parallax is too large relatively to neighbouring pixels. The resulting d_p parallaxes are stored separately in one D matrix having the same dimension as I_1 image.

In MATLAB we can setup several parameters to refine the result when applying the `disparityMap` function. These parameters are:

- Number of directions: 4 or 8.
- The smallest disparity.
- The biggest disparity.
- Block dimensions: how many pixels should be considered when calculating the energy function?
- Maxima disparity between the adjacent pixels (left/right neighbour).
- Uniqueness-ratio: limit, where the best energy value is the winner.

From these parameters the most important is to setup correctly the disparity range (smallest and largest disparity) according to Fig. 8. The disparity map as a result is shown on Fig. 9. The grey colours are representing the depth information.

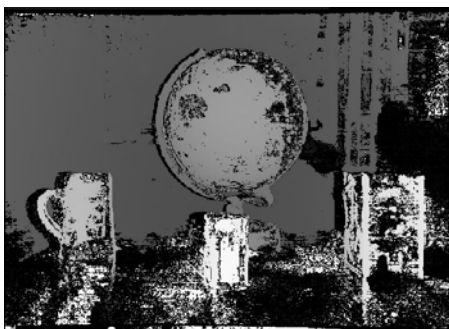


Fig. 9 Disparity map

On Fig. 9 we can notice disturbances and missing areas (in black colour). Because of the occlusion some parts are missing from our model. In MATLAB we can generate the point cloud with the help of the `reconstructScene` function. The visualized point cloud is demonstrated on Fig. 10, where the white areas mean holes in the model.



Fig. 8 Error ellipses on new points

V. CONCLUSIONS

As a summary we can say that the implementation of the SGM algorithm is greatly depending on the setup of parameters. The other important issue is that the SGM algorithm works only with stereo-pairs. In case of more stereo-pairs we need to integrate the separate models. The occlusion problem remains a problem. The experiments showed us that it is better if the parallaxes are not too large, so we need to take photos with small basis, but in this case the accuracy issues can play an important role.

REFERENCES

- [1] Hirschmüller H. – Semi-Global Matching – Motivation, Developments and Applications, in: Dieter Fritsch (Ed.): Photogrammetric Week'11, Stuttgart, Wichmann Verlag, Berlin, ISBN 978-3-87907-507-2, pp. 173-183., 2011
- [2] Luhman T., Robson S., Kyle S., Boehm J. – Close-Range Photogrammetry and 3D Imaging, Walter de Gruyter GmbH, Berlin/Boston, 2. kiadás, ISBN 978-3-11-030269-1, pp. 486-487., 2014
- [3] MathWorks, Inc. – Computer Vision System Toolbox, <https://www.mathworks.com/help/vision/index.html>, 2017

Supporting Historic Building Reconstruction by Laser Scanning

N. Rehany, T. Lovas

Department of Photogrammetry and Geoinformatics,
Budapest University of Technology and Economics, Budapest, Hungary
rehany.nikolett@epito.bme.hu, lovas.tamas@epito.bme.hu

Abstract—Due to reconstruction planning, complex building survey of the Royal Castle of Esztergom was requested by architects. Both horizontal and vertical layout of the castle as well as the wall surfaces are irregular, heterogeneous, the inner parts are dark, therefore laser scanning was chosen as data acquisition technique. Outside, especially at the higher walls, the laser scanning point cloud was supplemented with that of derived from Unmanned Aerial System (UAS) photogrammetry, where GNSS measurements enabled the georeferencing and additionally supported merging the point clouds.

Besides providing floor plans, sections and view for design purposes, historical architectural documentation of the current state of the building was also requested. The architects create detailed true-to-form drawings of the wall surfaces (by drawing all visible contour lines) extended by legends showing various attributes (e.g. any signs and damages on the wall surface). To support this work, orthoimages of the walls have been created that can be used and interpreted on-site by architects. Since the wall surfaces are very fragmented and irregular, moreover many of the rooms are dark, we had to find a solution that does not need optical imagery. Since laser scanning was used to capture the entire 3D geometry of the building, we created orthoviews derived from the point clouds. Reference planes were defined based on the architects' instructions then the relevant part of the point cloud has been cut and cleaned from the noise, and finally it has been projected to the reference plane. Owing to the intensity values, the contours (e.g. that of the stones, bricks) are nicely separated from other parts, hence architects can efficiently use it on-site to recognize and identify the particular wall segments. Additionally, virtual walk application (like Google Street View) was created that further supports both on-site architectural measurements and post-processing.

I. INTRODUCTION

The Department for History of Architecture and of Monuments of Budapest University of Technology and Economics ordered the detailed surveying of the Royal Castle of Esztergom due to reconstruction planning. The goal of the historical building research is to detect and reveal the circumstances of the construction and the changes occurred during the years with the analysis of the building, its structure, available sources/references (archives, plans, plots, documents, etc.). The architects apply the method called “Bauforschung;” the historical building investigation and surveying method developed and widely used in Germany [2], [7], [8], [10]. The damaging interventions on the building can be reduced to the necessary minimum with use of this procedure. The primary method of Bauforschung is the true-to-form

architectural survey which based on the precise surveying without any prior assumptions and searching for any regularity or parallelism [3], [4], [8]. The place, sizes and shape of the elements (corners of bricks and stones) are determined quite accurately and in details, in a coordinate system independent of the wall. Since the manual surveying drawing is made on-site, the architect can observe not only the visible but the tangible particularities. They register every important, supplementary information like stone damage, crack, weathering, replacements, wedges, visible marks of stone dressing tools (corrections, marks of masons) graphically and/or textually in the drawing. The true-to-form drawings are created in such quality and resolution that they could be used as important resource for further researches. The derived orthoimages are scanned, joined together, and then digitized; therefore enabling the vector-based drawing in CAD environment (usually AutoCAD or ArchiCAD).

The conventional Bauforschung surveying methods use accurately located horizontal and vertical reference strings on the walls; one architect uses them for the measurements, while others prepare markings on the wall on-site [5], [6], [7]. Other methods require less on-site measurement time; only a sketch is drawn by hand and photos taken from the wall texture and well identifiable points are measured e.g. by geodetic total stations. This technique require an additional surveying phase when supplementary information (e.g. stone damages; cracks, marks on the surface) are acquired on-site. On the next level a preliminary but accurate vector drawing based on on-site measurements was created, then the architects put the detailed geometry and attributes on the printed drawings on-site [4], [5], [6]. Photo-tacheometry was also used to support general architectural survey [9]. Further developments aimed at 3D modelling, attaching database that contains all relevant data and visualization in virtual reality [1].

It is important to note that the recently even more widely used laser scanners cannot substitute the process of the explanatory drawing method but can supplement and support it very well by surveying rooms with irregular shape or high wall segments. The photogrammetry itself is not suitable for the architectural surveying in such circumstances, but is able to provide a reasonable basis, too. It is difficult to distinguish (based only on imagery) between different building materials, cracks and shadows, and the vegetation and scaffolding could cover the important parts of the wall.

II. SURVEYING PLANNING

The first step of historical building research is the documentation of the current state in the most accurate

way. Our department, the Department of Photogrammetry and Geoinformatics, has been asked to survey the building and provide different derived products to support the work of the architects. Besides providing floor plans, sections and views for reconstruction planning, orthoviews of each wall in each room and the outside of the building was also requested.

In the practice, geodetic total stations are used for carrying out the true-to-form survey but due to the special circumstances we decided to use terrestrial laser scanner (TLS). The geometry of the building is complex, fragmented and irregular (narrow corridors; steep stairways; irregular, small rooms; complex, fancy doors), moreover, many of the rooms are dark without any natural light. The outside, especially the higher parts of the wall and the roof have been surveyed with Unmanned Aerial System (UAS) photogrammetry and a 3-dimensional point cloud have been generated from the photos to supplement the laser scanning surveying. In Fig. 1 the castle can be seen on images captured by UAS. GNSS (Global Navigation Satellite System) measurements have been performed to enable georeferencing and, in addition, to support merging the point clouds. Furthermore the surveying and research results become comparable with that of the previous surveyings.

The most challenging task was creating the orthoviews. Traditionally the photos are captured by high resolution DSLR camera, from near orthogonal positions, but this solution was not applicable in this case because of the dark rooms; lighting was not an option due to far power source, and extremely fragmented wall surfaces. Since laser scanning was used to capture the entire 3D geometry of the building, we created orthoviews derived from the point clouds. Since the darkness disabled the image capture of the scanner and hence deriving colored point clouds, we used the registered intensity values as thematic information. The bricks and stones appear decisively on the intensity colored point cloud, they could be separated from each other, which is very important from the architectural interpretation aspect.

III. SURVEYING AND DATA PROCESSING

The building has been surveyed with a Faro Focus 3D S120 terrestrial laser scanner and outside, where laser scanner could not be used due to the tall building and unfavorable incident angles, photos have been captured by



a GoPro Hero 3+ action camera mounted on a DJI Phantom quadcopter. The coordinates of the ground control points (in Fig. 1 red-white square tables on the ground) have been determined with GNSS technology along with the outdoor measurements. The number of scanner positions is depending on the complexity of the castle; as mentioned, many smaller rooms had to be surveyed then transformed into a common coordinate system. The scanning resolution was adjusted to the minimum geometric resolution required by the orthoviews. The castle was surveyed from 147 scanning positions, the joint point cloud contains 6 billion points, the size of the raw data is 31 GB. The mean range measurement error of the applied laser scanner is 2 mm and the minimum point spacing is 1.5 mm at 10 m distance. The applied resolution was 1/4 (6 mm spacing at 10 m) indoor and we selected 1/2 (3 mm spacing at 10 m) outdoor where long distances had to be captured. Due to long and fragmented traverse lines we used reference spheres as tie points to enable registering the point clouds to each other. Photos were taken by laser scanner only outdoor where the light conditions enabled, therefore colored point cloud is available only outside of the building. 456 photos have been captured by UAS that enabled to generate the point cloud of the building, however, it is much sparser than that of provided by TLS. Point cloud-to-point cloud registration technology (Iterative Closest Point - ICP) has been used to joint the separate point clouds. Due to the hardly manageable size of the exported joint point cloud each scan station's point cloud has been resampled with 1 cm point spacing. We used the scanner's own processing software (Faro Scene) and the free and open source CloudCompare to process the measurements and produce the final point cloud. The combined and resampled TLS and UAS point cloud consists of about 80 million points. This point density is more than enough for deriving floor plans and sections, however, it is not enough for orthoviews because important details could be lost with the resampling. To obtain high quality orthoviews, it is necessary to use each measured points, so we worked with the original density in this work phase.

For the orthoviews of walls, the parts of interest have been cut from the whole, original density point cloud and cleaned from noise (scaffoldings, tables, display panels, exhibits, visitors etc.). In general, the architects defined vertical planes as projection planes which were aligned to the relevant walls (Fig. 2). The cleaned point cloud



Figure 1. Photos about the castle captured by UAS

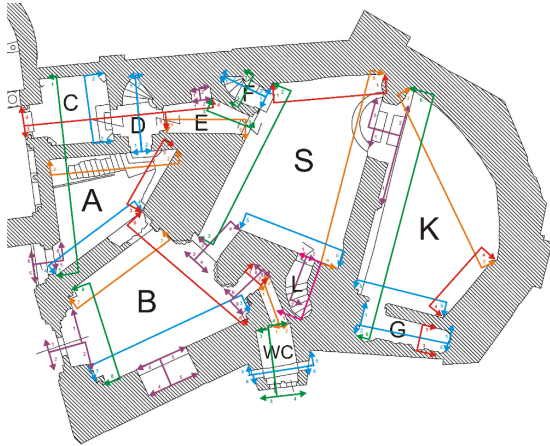


Figure 2. Floor plan of the castle's basic level with reference planes defined by architects

belonging to the particular wall has been rotated orthogonally to the defined projection plane of the same wall, then pixel size of the points has been set for optimal representation, finally the pictures have been exported in JPG format. The pictures have a 1-meter spacing grid overlay, therefore the orthoview of large walls could be cut into smaller pieces by architects and later merged together accurately. The vertical axis of the grid is aligned the Baltic Height System, while the vertical axis always begins at arbitrary place, somewhere left from the illustrated wall as it shown on Fig. 3. The surveyed part of the building contains more floors, so there are walls above each other; in these cases it was necessary to use exactly the same reference system and grid by generating orthoviews. In one of the basement rooms an excavated wall section can be found, and it was a special requisite to create an orthoview from the top of it and of the neighboring room but on the same picture. In this case horizontal reference plane has been defined and the

orthoview has been made perpendicular to it, from above. In another room the ceiling is consisted of vaults which is also important for architects to be documented, therefore an orthoview has been created about it; a horizontally reference plane has been used and the point cloud has been rotated from below so the ceiling has been viewed from below. Fig. 4 shows the mentioned ceiling orthoview and the corresponding floor plan; differing shapes are noticeable because the orthoview is from below, while floor plan is viewed from above. Altogether, besides these horizontal orthoimages, 20 orthoviews have been created from the outer walls, and 100 from the inner walls.

The created orthoviews have been recolored and faded by architects with an image processing software, then cut to suitable sizes, and finally plotted in scale 1:20. Architects have been drawn contours and supplementary information with pencil onto the plotted orthoviews on-site (see Fig. 5). The finished drawing have been scanned and joined together by the grid. Because of discoloring before printing, the point cloud is not visible on the drawing after scanning, however, it is visible on the printed paper so it can support the drawing on-site. The scanned drawings are the end products of this project, the lines have not been redrawn digitally.

Beyond the predefined products, virtual walk application (like Google Street View) has been created that allows to walk between the scanner positions and turn around virtually, so the entire scanner's field of view is visible through panorama pictures [11]. The position and orientation information were available for each scanner positions and panorama pictures were derived from the intensity colored point cloud (or from pictures where pictures were available) to create this application. This application enables the rapid check of the scene without the use of any point cloud manager software therefore it supports both on-site architectural measurements and post-processing.



Figure 3. Intensity colored orthoview of a wall with the overlay grid and labels (left) and detail from panorama picture about the same wall made by laser scanner (right)

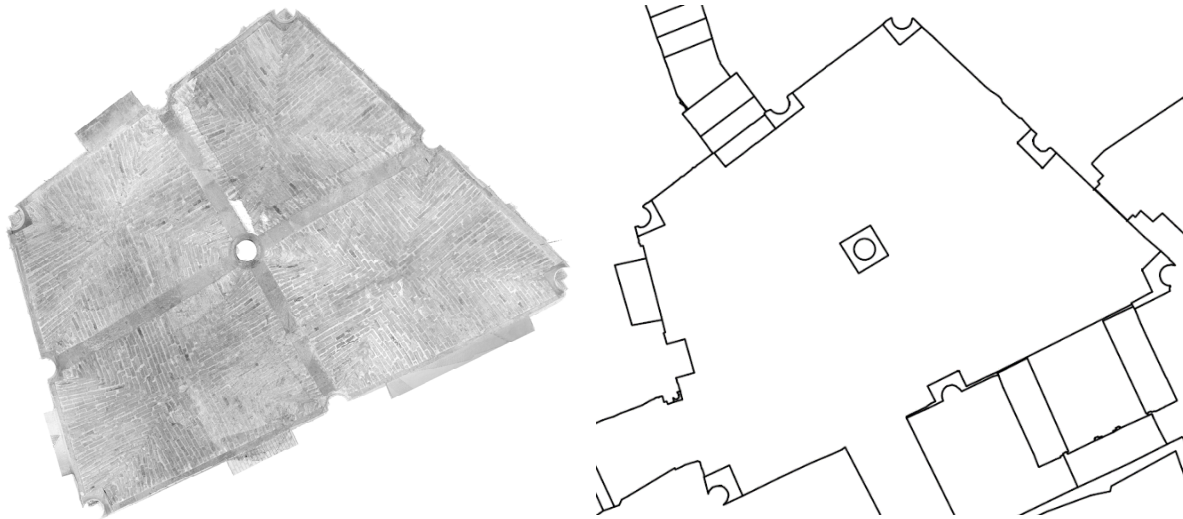


Figure 4. Bottom orthoview (left) and floor plan (right) about the vaulted room

IV. PROBLEMS OCCURRED

The first problem we had, that the two type of point clouds from different sources have different point density. The point cloud derived from UAS photos supplements the incomplete laser scanned point cloud but it is not dense enough for true-to-form survey; contours of stones and bricks could not be differentiated (see Fig. 6 (left)). Table 1 shows the average point density of the merged point cloud both inside and outside of the building.

The point cloud derived from the UAS measurements is only colored, so true colored point cloud was used (instead of intensity based one) for the merged dataset. Due to the two different types of cameras (GoPro and the built-in camera of the scanner), changing lighting conditions (sun, clouds) and shadows, the walls have a bit different colors in the pictures.

TABLE I.
POINT DENSITY OF THE MERGED POINT CLOUD INSIDE AND OUTSIDE OF THE BUILDING

		Points / m ²
Inside	Minimum	154 414
	Average	2 662 554
	Maximum	10 622 378
Outside	Minimum	6 891
	Average	114 343
	Maximum	2 002 183

Mostly at the higher parts of walls, due to lower point density, the contours of the bricks are not unambiguously selectable as nicely as at the bottom of the walls, as shown in Fig. 6.

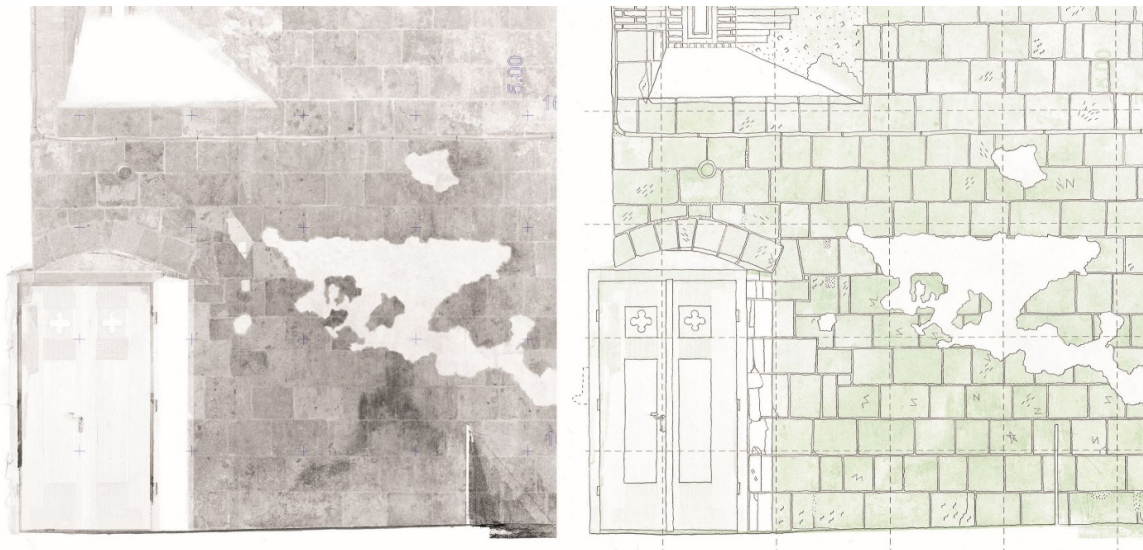


Figure 5. Orthoview of wall section (left) and the result of true-to-form survey based on in, made by architects (right)



Figure 6. Merged (TLS + UAS) point cloud on orthoviews about a higher wall part and roof (left) and bottom of a wall (right) outside

Compared to image pixel intensity values, the laser intensity values carry more information on the material of the surface that is very useful in such situations when elements with similar color but different material are to be separate. In Fig. 7 the difference between the mentioned intensity values can be observed.

A room has been scanned with color i.e. the scanner also took photos after scanning, resulting a colored point cloud. Unfortunately, the colored point cloud cannot be used, much more light would be needed even in the best

illuminated room. Fig. 3 (right) represents a part of panorama photo of this room. Therefore only intensity colored point clouds have been used inside the castle but this type of data also has its limit. When the scanner position was closer to the wall than the minimum distance recommended by manufacturer, concentric, dark part appeared on the point cloud even on flat, uniform surfaces. As it can be seen on Fig. 8, this effect could not be avoided in some places like on the narrow stairway.



Figure 7. True colored (left), intensity colored (middle) point cloud and a photo (right) of the same wall part

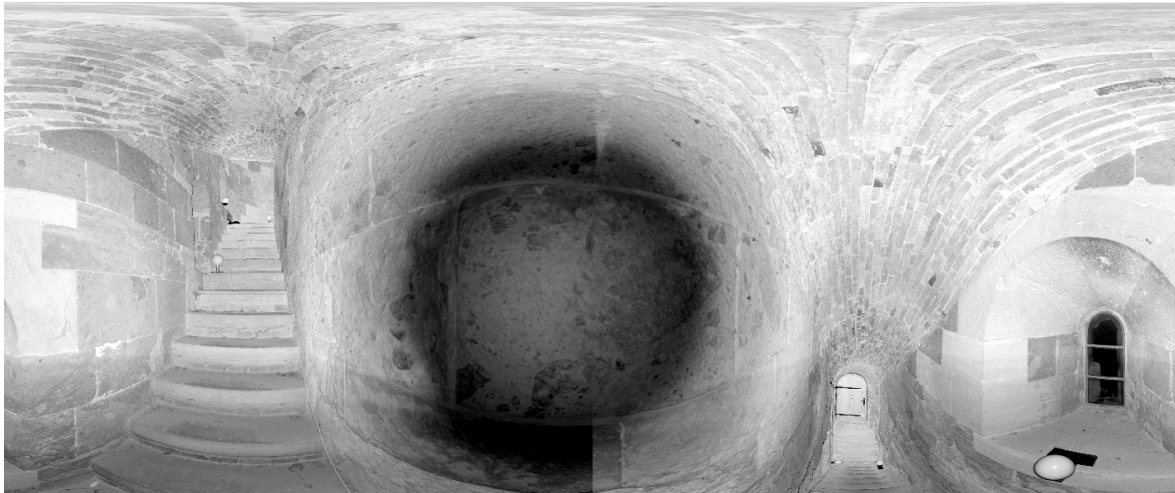


Figure 8. Concentric, dark part on panorama picture derived from intensity point cloud of a station on narrow stairway

V. CONCLUSIONS

A workflow has been developed for supporting historical building research. It can be efficiently used in a place, where it is not possible to create the requested products with traditional optical sensors. The orthoviews derived from intensity colored laser scanned point cloud can support on-site architectural work in case of dark rooms and fragmented walls, too. Due to the comprehensive 3D surveying, the reference planes can be defined in any directions and therefore orthoviews have been generated also about ceiling and floor. Solutions have been developed and applied by adaptive parameterization of laser scanning and using data fusion to overcome the occurred problems.

VI. FUTURE PERSPECTIVES

We plan to develop a workflow which allows architects to bring the orthoviews in digitally format to the field and creating vector-drawings on-site. Hereby plotting the orthoviews, drawing with pencil, scanning the pictures and redrawing could be substituted. Our goal is involving a GIS software that would enable assigning important notes, attributes, or even photos to the drawing entities. For example, if the architects would draw on a tablet and take photos with the same device, the photos could be assigned to each individual elements (e.g. one stone in the wall) during drawing. The current workflow used by architects would allow to join the digitized, vector data and the notes, important attributes, hereby making a geodatabase of the true-to-form survey. In practice, this time-consuming postproduction usually does not occur, however, this could support understanding the results of the research to the wide public. The geodatabase would also support visualizing the result of the research, e.g. by generating thematic maps of the walls.

REFERENCES

- [1] P. Drap et al., "Going to Shawbak (Jordan) and getting the DATA back: toward a 3D GIS dedicated to medieval archaeology," *3D Virtual Reconstruction and Visualization of Complex Architectures*, vol. XXXVIII-5/W1, February 2009 [3rd ISPRS International Workshop 3D-ARCH, 2009].
- [2] G. U. Grossmann, *Einführung in die Historische Bauforschung*. Darmstadt: Wissenschaftliche Buchgesellschaft, 1993.
- [3] Gy. Hajnoczi, "Mümlékfelmérés," *Az Építőipari Műszaki Egyetem Tudományos Közleményei*, vol. 1/6, Budapest, 1956.
- [4] B. Halmos, K. Marotzy, "The adaptations of the true-to-form survey method," in *Periodica Polytechnica Architecture*, vol. 2010/2, pp. 9-17, 2010.
- [5] B. Halmos, K. Marotzy, "Észrevételek a gyulafehérvári Szent Mihály-székesegyház szentélyének felmérése nyomán," in *A gyulafehérvári székesegyház főszentélye*, Sz. Papp, Ed. Budapest: Teleki László Alapítvány, 2012, pp. 43-58.
- [6] B. Halmos, K. Marotzy, G. D. Nagy, "A gyulafehérvári székesegyház déli tornya," in *Dolgozatok az Erdélyi Múzeum Érem- és Régiségtárából*, vol. XVI-XVII, I. Bajusz, T. Emodi, E. Benko, A. Kovacs and A. Laszlo, Eds. Kolozsvár: Erdélyi Múzeum-Egyesület, 2013, pp. 217-239.
- [7] J. Krahling, B. Halmos, J. Cs. Fekete, "A fertődi marionettszínház új értelmezése - az épületkutatás ("Bauforschung") és alakmű felmérés mint kutatási módszer alkalmazásával," (in Hungarian), *Építés-Építészettudomány*, vol. 34 (1-2), pp. 5-55, 2006.
- [8] J. Krahling, "Épületkutatás és építészettörténet - néhány újkori és 19. századi épülettípus kutatásának módszertana és eredményei," habilitation theses, *Építés-Építészettudomány* vol. 45 (3-4), pp. 341-364, 2017.
- [9] M. Scherer, "Architectural Surveying and Visualization Using "Photo-Tachometry"," *From Pharaohs to Geoinformatics*, Cairo, Egypt, April 2005 [FIG Working Week 2005 and GSDI-8].
- [10] M. Schuller, "Building Archeology," in *Monuments and Sites*, vol. VII, München-Paris: ICOMOS, 2002.
- [11] http://web.fmt.bme.hu/fmtpontfelho/StreetView_Esztergom



SUPPORTED THROUGH THE NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

Some Ramsey Type Problems in Mathematical Competitions

Borbély József

Óbudai Egyetem, Székesfehérvár, Hungary

borbely.jozsef@amk.uni-obuda.hu

Abstract— In this paper we will discuss some Ramsey-type problems, that occurred in mathematical competitions. We will also emphasize the non-elementary background of the problems.

I. INTRODUCTION

Ramsey-type problems form a crucial part of combinatorial research. In every branch of mathematics there are theorems, that can be solved using ideas of Ramsey-type theorems.

The Ramsey theory was baptised after the British mathematician and philosopher Ramsey. The objects of Ramsey's original theorem are graphs, but the research was essentially extended implementing similar ideas in other mathematical structures.

Ramsey-type problems can be formulated in the following way: if we take a mathematical structure large enough and order its elements in classes, then there must be a class containing a “regular” subclass.

We have to emphasize the difference between Ramsey-type theorems and density theorems. A density theorem simply states that by choosing a relatively large subset of a given set, the chosen subset must contain (by its cardinality) a “regular” structure large enough.

II. SOME CLASSICAL RAMSEY-TYPE AND DENSITY THEOREMS

Eventually the first Ramsey-type problem appeared in an old Kürschák competition:

Exercise 1: *In a group of six peoples there are three that pairwise know each other or three that pairwise do not know each other.*

Really, in Exercise 1 we ordered the edges of a simple complete graph with six vertices into two classes (acquaintance and non-acquaintance). We can say, that the edges were colored by two colors, and so we can reformulate the statement of Problem 1:

Exercise 1 reformulated: *If we color all the edges of a simple complete graph with six vertices by red and blue, then the graph must contain a red or a blue triangle.*

From now on, we will formulate our Ramsey-type theorems with the coloring terminology (that symbolizes the ordering into classes). We will speak about s -coloring, if we color the edges (or the objects) by s colors.

The general Ramsey theorem was proved by Ramsey, and generalizes the result of Exercise 1.

Theorem 1 (Ramsey's theorem): *For every positive integer s and positive integers n_1, n_2, \dots, n_s , there is a smallest integer $R=R(n_1, n_2, \dots, n_s)$, such that for every s -coloring of the edges of a simple complete graph with at least R vertices we can find a complete*

subgraph with n_1 vertices whose edges were colored by color 1, or a complete subgraph with n_2 vertices whose edges were colored by color 2, ..., or a complete subgraph with n_s vertices whose edges were colored by color s .

Generally it is hard to determine the exact values of the above defined so-called Ramsey numbers $R(n_1, n_2, \dots, n_s)$. The result of Exercise 1 means that $R(3,3) \leq 6$ (here $s=2$ and $n_1=n_2=3$). It is easy to show that $R(3,3)=6$.

Now we will present a classical density theorem concerning graphs.

Theorem 2 (Turán's theorem): Let k be a positive integer. If a simple graph with n vertices has more than $\left(1 - \frac{1}{k}\right) \frac{n^2}{2}$ edges, then it contains a complete subgraph with $(k+1)$ vertices and this bound is the best possible.

We can easily find Ramsey type results by combining Ramsey's and Turán's theorems. The following exercise was posed in "Matematikai Lapok" in Erdély:

Exercise 2: For every 2-coloring of the edges of a simple graph with $5n$ vertices and with $10n^2+1$ edges the graph must contain a monochromatic triangle.

Exercise 2 can be proved easily in the following way: by Turán's theorem every simple graph with $5n$ vertices and with $10n^2+1$ edges contains a complete subgraph with 6 vertices. Using the fact that $R(3,3)=6$, the statement follows.

But we have to take into consideration, that the solution of Exercise 2 was easy only because we knew Turán's theorem. But Exercise 2 is not a classical Ramsey type problem in the strict sense, because the graph in which we sought for a complete subgraph with 6 vertices was not complete.

The statement of Exercise 2 can be generalized easily in the following manner:

Theorem 3: Let n_1, n_2, \dots, n_s be positive integers. Then there is a positive integer m such that for every integer n greater than m the following holds:

for-every s -coloring of the edges of a simple graph with n vertices and more than

$\left(1 - \frac{1}{R(n_1, n_2, \dots, n_s) - 1}\right) \frac{n^2}{2}$ edges one can find a complete subgraph with n_1 vertices whose edges were colored by

color 1, or a complete subgraph with n_2 vertices whose edges were colored by color 2, ..., or a complete subgraph with n_s vertices whose edges were colored by color s .

Now we will present some Ramsey type problems that concern several branches of mathematics.

III. SOME RAMSEY- TYPE THEOREMS AND PROBLEMS IN GEOMETRY

We will open this section with two elementary exercises:

Exercise 3: For every 2-coloring of the points of the plane one can find a regular triangle such that its vertices are colored by the same color.

Exercise 4: For every 2-coloring of the points of the plane one can find a triangle such that its vertices and its centroid are colored by the same color.

The common origin of Exercise 3 and Exercise 4 is the so-called Gallai-theorem:

Theorem 4 (Gallai): Let m and s be a positive integer and let $V = \{v_1, v_2, \dots, v_f\}$ be a finite set of points in the m -dimensional plane. Then for every s -coloring of the m -dimensional plane one can find a monochromatic set of points homothetic to V , i.e. there is a real number c and an m -dimensional vector w such that the points of the set $cV + m = \{cv_1 + w, cv_2 + w, \dots, cv_f + w\}$ have the same color.

The statements formulated in Exercises 3 and 4 are easy consequences of Gallai's theorem. But by changing the conditions, we can get Ramsey type results, that are not direct consequences of Gallai's theorem.

The following exercise can be solved with the pigeonhole principle:

Exercise 5: Let m, d and s be positive integers, where $s \leq d \leq m$. Then for every s -coloring of the points of the m -dimensional plane one can find two monochromatic points which are one unit apart.

Really, if we consider a regular tetrahedron with edges of length one in the m -dimensional space, then this polygon has $(m+1)$ vertices, and by the pigeonhole principle it must have two vertices with the same color.

The following exercise was posed in the national competition of Iran:

Exercise 6: *Let ABC a given triangle. If one colors the points of the plane by red and green, then one can find two red points which are one unit apart or one can find three green points which are the vertices of a triangle congruent to ABC .*

Exercise 5 and 6 are interesting because they ask for congruences, thus we cannot use Gallai's theorem.

If we take the problems posed in Exercise 5 and 6 into consideration, we can ask for a more general result.

Borbély has the following conjecture concerning these type of problems:

Conjecture (Borbély): *There is a polynomial p , such that for every finite sets of points V_1, V_2, \dots, V_s with the property $|V_1| + |V_2| + \dots + |V_s| \leq p(d)$, then the following holds:*

for every s -coloring of the points of the d -dimensional space, there is a set of points congruent to V_1 in which the points are colored by color 1, or there is a set of points congruent to V_2 in which the points are colored by color 2, ..., or there is a set of points congruent to V_s in which the points are colored by color s .

Another way to generalize Gallai's theorem is to take other sets in the place of the whole m -dimensional plane.

The following exercise was posed in the Hungarian national competition OKTV:

Exercise 7: *By coloring the lattice points of the plane by six colors, one can find a rectangle such that its sides are parallel to the axes and its vertices have the same color.*

Exercise 7 has also an elementary solution using pigeonhole principle. But for lattice points one can generalize

Gallai's theorem in the following manner:

Theorem 5 (Gallai-Witt): *Let m and s be a positive integer and let $V = \{v_1, v_2, \dots, v_f\}$ be a finite set of the lattice points in the m -dimensional plane. Then for every s -coloring of the lattice points of the m -dimensional plane one can find a monochromatic set of points homothetic to V , i.e. there is an integer c and an m -dimensional vector w with integer coordinates such that the points of the set $cV + m = \{cv_1 + w, cv_2 + w, \dots, cv_f + w\}$ have the same color.*

It is straightforward that the Gallai-Witt theorem implies the statement of Exercise 7.

IV. SOME RAMSEY-TYPE THEOREMS AND PROBLEMS IN ALGEBRA

We open this section with some exercises. The following exercises were originally posed in Kömal:

Exercise 8:

a, Prove that among any six irrational numbers there are three such that the sum of every two of them is irrational.

b, Prove that among any five irrational numbers there are three such that the sum of every two of them is irrational.

The statement of part a, is a direct consequence of the fact, that $R(3,3)=6$. Part b, is surprising and its validity is due to the special properties of rational and irrational numbers. In this sense Part b, can be considered as a special Ramsey result.

We will go much further and will prove the following theorem (once released in Kömal among the hard problems):

Theorem 6:

Among any $(2n-1)$ irrational numbers there are n (say x_1, x_2, \dots, x_n) with the following property: if r_1, r_2, \dots, r_n are nonnegative real numbers such that not all of them are zeros, then

$$r_1x_1 + r_2x_2 + \dots + r_nx_n \text{ is irrational.}$$

Clearly Exercise 8 is a very special case of Theorem 6.

V. SOME RAMSEY TYPE THEOREMS AND PROBLEMS IN NUMBER THEORY

Applying the Gallai-Witt-theorem in one dimension we get the following celebrated result of Van der Waerden:

Theorem 7 (Van der Waerden): *For every k and s positive integers there exists a smallest integer*

$W(s,k)=W$, such that for every integer not less than W the following holds:

for every s -coloring of the elements of the set $\{1,2,\dots, n\}$ one can find a k -term monochromatic arithmetic progression in the set.

It is hard to exactly determine the above defined so-called Van der Waerden numbers $W(s,k)$.

A famous elementary problem that occurred in OKTV as a Van der Waerden type result is the following:

Exercise 9: The elements of the set $\{1,2,\dots, 1986\}$ can be colored by two colors in such a way that one cannot find a 18-term monochromatic arithmetic progression.

Clearly in Exercise 9 one has to prove that $W(2, 18) > 1986$. One can give a direct construction by coloring the numbers in $\{1,2,\dots,1986\}$ that are divisible by exactly one of the numbers 7 and 17 by red, and the other numbers by blue. It is not hard to show that this construction works.

By using Berlekamp's theorem we can give a much more general and stronger estimate:

Theorem 8 (Berlekamp): For every prime number p holds the inequality $p \cdot 2^p \leq W(2, p+1)$.

Applying Theorem 8 for $p=17$ we get that $17 \cdot 2^{17} \leq W(2, 18)$, which is much stronger than the result of Exercise 9.

ACKNOWLEDGMENT

We want to express our gratitude towards the Hungarian Ministry of the Human Capacities and the Óbudai Egyetem for having generously granted the research work.

REFERENCES

- [1] Graham, Rotschild and Spencer, *Ramsey Theory*, Wiley, 1980.
- [2] Alexander Soifer, *The Mathematical Coloring Book*, Springer, 2009.



Some Linear Algebraic Problems

Borbély József, Csala-Takács Éva

Óbudai Egyetem, Székesfehérvár, Hungary

borbely.jozsef@amk.uni-obuda.hu

csala.takacs@amk.uni-obuda.hu

Abstract:

In this paper we will present some problems which are related to topics in linear algebra.

Definition: Let V be a vector space over the field F . The vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ form a linearly independent system of vectors, if for the elements t_1, t_2, \dots, t_k of F the equality $t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \dots + t_k\mathbf{v}_k = \mathbf{0}$ implies that $t_1 = t_2 = \dots = t_k = 0$. If a system of vectors is not linearly independent, we call it linearly dependent.

Introduction

Using linear algebra is an efficient way to solve mathematical problems. We will demonstrate the force of linear algebra through various examples.

The presented problems are elementary, but can be considered as hard problems. Using the presented methods one can easily generalize the problems and pose new problems.

Related to linear independence, we will prove a result, which appeared in Kömal among the hard problems. The problem's object is rationality and irrationality of numbers generated by linear combinations.

The solution does not require non-elementary ideas, but that does not mean, that the problem was easy. If someone wants to solve this problem, then one has to know the standard ideas of linear algebra.

Linear dependence and independence

One of the most important terms in linear algebra is the one of linear independence.

Theorem 1:

Among any $(2n-1)$ irrational numbers there are n (say x_1, x_2, \dots, x_n) with the following property: if r_1, r_2, \dots, r_n are nonnegative real numbers such that not all of them are zero, then

$r_1x_1 + r_2x_2 + \dots + r_nx_n$ is irrational.

Proof:

We will define linear independence and dependence before the investigation of our first problem.

We will prove the statement of the theorem by induction on n . If $n=1$, then the statement holds.

Let us assume that we proved the statement for $n < N$. Let $x_1, x_2, \dots, x_{2N-1}$ be irrational numbers. Let us assume to the contrary, that the statement of the theorem does not hold for these numbers.

Using the induction hypothesis we can find $(N-1)$ among these $(2N-1)$ numbers (say x_1, x_2, \dots, x_{N-1}), which fulfill the conditions. Due to the indirect assumption one cannot choose an N -th number from the set $\{x_N, x_{N+1}, \dots, x_{2N-1}\}$ such that the prescribed condition holds.

There are nonnegative rational numbers $b_N, b_{N+1}, \dots, b_{2N-1}$ such that not all of them are zeros and the number

$$b_N x_N + b_{N+1} x_{N+1} + \dots + b_{2N-1} x_{2N-1} \text{ is rational.}$$

Let us consider the N -element sets $H_j = \{x_1, x_2, \dots, x_{N-1}, x_{N-1+j}\}$, where $j=1, 2, \dots, N$.

Due to the indirect assumption for every $j=1, 2, \dots, N$ there are nonnegative rational numbers

$a_{j,1}, a_{j,2}, \dots, a_{j,N}$, such that not all of them are zeros and the number

$$a_{j,1} x_1 + a_{j,2} x_2 + \dots + a_{j,N-1} x_{N-1} + a_{j,N} x_{N-1+j} \text{ is rational.}$$

By multiplying by appropriate rational numbers one can attain that $a_{j,N} = b_{N-1+j}$ holds for every $j=1, 2, \dots, N$.

Now let us sum up the last N linear combinations, let us label this sum by S .

The sum S must be a rational number, because all of the linear combinations were rational.

But this sum S can be presented as the sum of the rational number $b_N x_N + b_{N+1} x_{N+1} + \dots + b_{2N-1} x_{2N-1}$ and a

linear combination of the numbers x_1, x_2, \dots, x_{N-1} , where all coefficients are rational and not all of the coefficients are zeros. But this implies that S must be irrational, which is a contradiction.

Remark 1:

in this problem one cannot speak about linear dependence and independence in the classical sense, because the investigation was restricted only to nonnegative rational

coefficients. We could speak about a “quasi-(in)dependence” in this case.

Remark 2:

the bound in the statement is the best possible, because by choosing

$x_1 = x_2 = \dots = x_n = e, x_{n+1} = x_{n+2} = \dots = x_{2n-1} = -e$ one can see that we cannot choose more than n “quasi-independent“

numbers in the desired manner.

Some plane geometry

We will show the efficiency of linear algebraic methods proving a result in plane geometry.

The following problem was posed in the journal “Matematika Tanítása”. The presented solution was published in the journal and was given by Borbély.

Theorem 2:

Let S be the circumscribed sphere of a regular tetrahedron whose edges have length l . If A_1, A_2, A_3, A_4 are points on the surface of S such that $|A_i A_j| < l$ for every i and j , then these four points lie on a hemisphere of S .

Proof:

Let O be the center of S and let \mathbf{x} denote the vector \overrightarrow{OX} . Let $ABCD$ a regular tetrahedron inscribed in S , $AB=BC=CD=DA=AC=BD=1, OA=OB=OC=OD=R$. Then

$$\mathbf{a} + \mathbf{b} + \mathbf{c} + \mathbf{d} = \mathbf{0}, \text{ thus } 0 = (\mathbf{a} + \mathbf{b} + \mathbf{c} + \mathbf{d})^2 = 4R^2 + 12R^2 \cos q,$$

where q denotes the angle AOB .

$$\text{Thus we get that } \cos q = -\frac{1}{3}.$$

We will prove that the tetrahedron $A_1 A_2 A_3 A_4$ does not contain the point O in its interior. From that follows the statement of the theorem immediately.

Let us assume to the contrary that the tetrahedron $A_1 A_2 A_3 A_4$ does not contain the point O in its interior.

With our notation this means that there are positive numbers t_1, t_2, t_3, t_4 such that

$$t_1 \mathbf{a}_1 + t_2 \mathbf{a}_2 + t_3 \mathbf{a}_3 + t_4 \mathbf{a}_4 = \mathbf{0}.$$

The condition $|A_i A_j| < 1$ can be equivalently reformulated as $\mathbf{a}_i \mathbf{a}_j > -\frac{1}{3}$.

Thus we have (by using that the t_j 's are positive)

$$0 = (t_1 \mathbf{a}_1 + t_2 \mathbf{a}_2 + t_3 \mathbf{a}_3 + t_4 \mathbf{a}_4)^2 =$$

$$R^2(t_1^2 + t_2^2 + t_3^2 + t_4^2) + 2(t_1 t_2 \mathbf{a}_1 \mathbf{a}_2 + t_1 t_3 \mathbf{a}_1 \mathbf{a}_3 + t_1 t_4 \mathbf{a}_1 \mathbf{a}_4 + t_2 t_3 \mathbf{a}_2 \mathbf{a}_3 + t_2 t_4 \mathbf{a}_2 \mathbf{a}_4 + t_3 t_4 \mathbf{a}_3 \mathbf{a}_4) >$$

$$> R^2(t_1^2 + t_2^2 + t_3^2 + t_4^2) + 2\left(-\frac{1}{3}R^2\right)(t_1 t_2 + t_1 t_3 + t_1 t_4 + t_2 t_3 + t_2 t_4 + t_3 t_4) =$$

$$= \left(\frac{1}{3}R^2\right) \left((t_1 - t_2)^2 + (t_1 - t_3)^2 + (t_1 - t_4)^2 + (t_2 - t_3)^2 + (t_2 - t_4)^2 + (t_3 - t_4)^2 \right) \geq 0,$$

which is a contradiction.

Proving an inequality

The following problem was posed in the journal "Matematika Tanítása". The presented solution was published in the journal and was given by Borbély.

Theorem 3:

If A, B and C are the angles of an acute angled triangle and x, y, z are positive numbers, then the inequality

$$x^2 \operatorname{tg} A + y^2 \operatorname{tg} B + z^2 \operatorname{tg} C \geq 2xyz \left(\frac{\sin A}{x} + \frac{\sin B}{y} + \frac{\sin C}{z} \right) \text{ holds.}$$

Proof:

We will use the following result of Sylvester as a lemma:

Lemma 1:

A Hermitian matrix is positive semidefinite if and only if all its principal minors are nonnegative.

Now let us reformulate the problem as

$$(x, y, z) \begin{pmatrix} \operatorname{tg} A & -\sin C & -\sin B \\ -\sin C & \operatorname{tg} B & -\sin A \\ -\sin B & -\sin A & \operatorname{tg} C \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \geq 0.$$

We will prove that the matrix

$$\begin{pmatrix} \operatorname{tg} A & -\sin C & -\sin B \\ -\sin C & \operatorname{tg} B & -\sin A \\ -\sin B & -\sin A & \operatorname{tg} C \end{pmatrix} \text{ is positive semidefinite.}$$

$\operatorname{tg} A > 0$, because A is an acute angle.

To prove that the second and the third principal minors are not negative, we will use the following lemma:

Lemma 2:

If A, B, C are the angles of a triangle, then

$$(1) \operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C = \operatorname{tg} A \operatorname{tg} B \operatorname{tg} C$$

$$(2) \sin 2A + \sin 2B + \sin 2C = 4(\sin A)(\sin B)(\sin C)$$

Using (1) we can write the second principal minor as

$$\operatorname{tg} A \operatorname{tg} B - \sin^2 C = \frac{\operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C}{\operatorname{tg} C} - \sin^2 C = \frac{1}{\operatorname{tg} C} (\operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C - \operatorname{tg} C \cdot \sin^2 C) =$$

$$= \frac{1}{\operatorname{tg} C} (\operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C \cdot \cos^2 C) > 0.$$

Using (1) and (2) the third principal minor can be written as

$$\operatorname{tg} A \cdot \operatorname{tg} B \cdot \operatorname{tg} C - \sin A \cdot \sin B \cdot \sin C - \sin A \cdot \sin B \cdot \sin C - \sin^2 B \operatorname{tg} B - \sin^2 A \operatorname{tg} A - \sin^2 C \cdot \operatorname{tg} C =$$

$$= \operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C - 2 \sin A \cdot \sin B \cdot \sin C - \sin^2 B \operatorname{tg} B - \sin^2 A \operatorname{tg} A - \sin^2 C \cdot \operatorname{tg} C =$$

$$= \cos^2 A \cdot \operatorname{tg} A + \cos^2 B \cdot \operatorname{tg} B + \cos^2 C \cdot \operatorname{tg} C - 2 \sin A \cdot \sin B \cdot \sin C =$$

$$= \frac{1}{2}(\sin 2A + \sin 2B + \sin 2C) - 2 \sin A \cdot \sin B \cdot \sin C = 0.$$

Thus the matrix is positive semidefinite and the statement of the theorem follows.

Acknowledgements

We want to express our gratitude towards the Hungarian Ministry of the Human Capacities and the Óbudai Egyetem for having generously granted the research work.

References

- [1] A matematika tanítása, issue 2011/3
- [2] A matematika tanítása, issue 2012/3
- [3] Kömal, issue 2003/5
- [4] Róbert Freud, *Lineáris Algebra*, ELTE Eötvös Kiadó, 2004

Business process sustainability in cloud based SaaS environments

István Orosz*

* Alba Regia Technical Faculty, Óbuda University, Székesfehérvár, Hungary
 orosz.istvan@amk.uni-obuda.hu

Abstract— A new software abstraction layer above the implementation layers was created in the SaaS based cloud technology and therefore heavily changed the way how Enterprise Resource Planning systems (ERP) are maintained and implemented over the hardware related platforms. The decades old release-by-release maintenance methodology, which was governed by version changes (pre-alpha to gold release) was substituted by a continuous release management. The Software as a Service (SaaS) model in the cloud which core business logic is implemented as a service offer above the low-level hardware close implementation layer. This architecture can offer software products which have longer lifecycle, because it is detached from the constantly changing physical implementation layer. This kind of a sudden change of technology is present in the latest IT architecture, the presence of an additional abstraction layer seems logical. The root cause is that the basic business processes are not changing so rapidly, so they can remain untouched under the hood. The SaaS type life cycle management means that the heavily technology independent part are not describing the business processes anymore. Previous lifecycle implementations from the assessment phase to the post go-live and support phase dealt the business logic as one entity with its implementation. That means, that the question of code reusability has a different role as in the standard on premise model. This paper introduces a new method of encapsulating and identifying the software parts, which can be later reused in a cloud SaaS environment.

Keywords: ERP, Model Driven Architecture Code sustainability, SaaS, BPR

I. INTRODUCTION

The total cost of ownership of an Enterprise Resource planning system is a very complex scenario to calculate. The currently used agile ones makes the whole calculation even more complex, because the agile methodology works in development circles, and there is now predefined number of the circles. The implementation period can be divided into several phases, such as the following [1][2]:

- Diagnostic
- Analysis
- Design
- Development
- Deployment

- Operation

One of the most expensive part is the development phase, this part is usually done by special teams. Therefore, it is a vital question, when it comes to a version upgrade, is there a way to determine which parts of the code can be reused without major modifications?

Previously, the on premise operational model offered an easy answer to this question. Between version upgrades, the operation model was not changed, therefore the code upgrade could take place in the same abstraction level. This standard model used the traditional client-server like architecture, where the following levels could store code elements:

- Client (thick client in most cases)
- Application server
- SQL database

This operational model was totally changed, when the ERP systems occupied their place in the cloud. The physical infrastructure moved from the datacenter to the cloud infrastructure, which means that the whole infrastructure operation is not in the scope anymore. On the other hand, bandwidth become a critical point of operation:

	On Premise	Cloud
Security	No dedicated internal IT group for security	Dedicated IT resources are needed to handle security internally
Costs	Cost are in correlation with the required resources, can be changed rapidly	Cost are in correlation with the size of the infrastructure
Maintenance	No dedicated IT support group	Dedicated IT support group with the company

Table.1. On premise and cloud based operational models comparison

Although the differences could be small for the first sight, the Platform as a Service and Software as a Service operational models make it a little complex, by implementing a new abstraction layer for the business core logic, which makes the biggest difference.

The cloud based Platform as a Service (PaaS) enables third party software solution service delivery, and covers the complexity of the implementation layer below. This cloud implementation layer means the whole middleware and datacenter technical solution. So, it offers the platform services, while hiding the whole architecture implementation from the solution in the cloud.

Software as a Service (SaaS) model offer a more sophisticated operational mode, with the following key features [3]:

- Data and the Software solution is stored in a common place, which is widely known as the cloud solution.
- Clients access the software solution via thin client, and are charged on usage base – not on license base
- Infrastructure implementation is present on the vendor, and the resource need can be reconfigured without any hardware side changes on the customer.
- Customers use the latest version of the software, provided on-demand, and do not have to bother with version upgrade.

This article focuses on the SaaS operating model, and describes a method, which can help identifying the reusable software components.

II. SOFTWARE REUSABILITY IN CLOUD BASED SAAS MODEL

This paper focuses on the code part reusability of a cloud based software solution in a Software as a Service operation model, as shown in Fig.1.

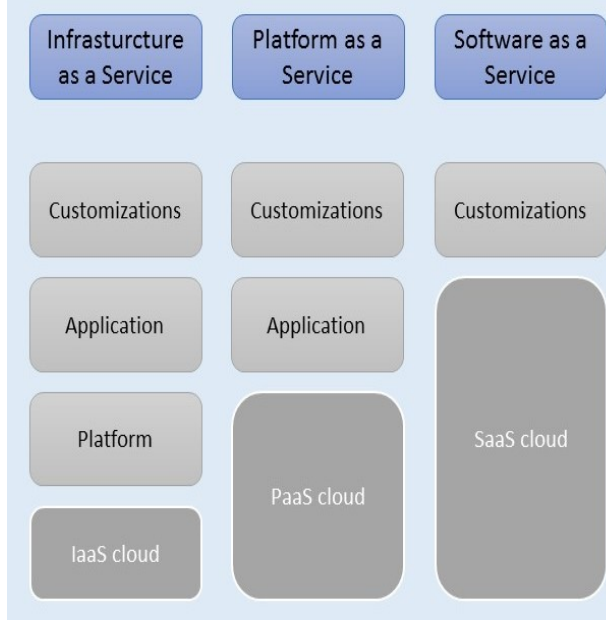


Fig.1. ERP cloud models [7]

This operating model presents a new abstraction layer, which means that software products has to adapt to the new requirements. Therefore, some layers of the original on-premise version have to be rewritten. The basic task is how to minimize the amount of code, which needs modification [13]?

The actual software product, which is the center of the interest is the Microsoft Dynamics 365 cloud version, which uses Microsoft Azure cloud as Platforms as a Service (PaaS), which offers the solution as a service all in one.

PaaS platforms makes possible for software solutions build and deliver services suitable to the SaaS model, and hides the highly complicated underlying middleware and the whole datacenter architecture [5].

- Infrastructure as a Service: provides a low level of hosting services, the service level is responsible for the virtualization, servers, DB storage, and network. It is the first cloud based operation level over the datacenter on-demand solution.
- Platform as a Service: one level above the IaaS, this service level is responsible for the operating system(s), middleware interface(s) and the runtime software modules. PaaS offers the control over the application and its data for the customers.
- Software as a Service: at this level, the customer(s) has to take care only modelling the core business logic over the software solution, and use the IaaS cloud based infrastructure as a whole implementation service. The total inverse of the on-demand datacenter model.

A. Model Driven Architecture (MDA)

MDA development approach is based on models, which acts as the foundation of design, development and the operation lifecycle. Separates the core business logic from the technical implementation. This will be the key point, which will act as the starting point of distinguishing the code part which can be refactored.

MDA uses the following three categories:

- Computation Independent Model, contains the business domain model
- Platform Independent Model, describes the system functionality in implementation method independent form
- Platform Specific Model, system specification according to the implementation technology

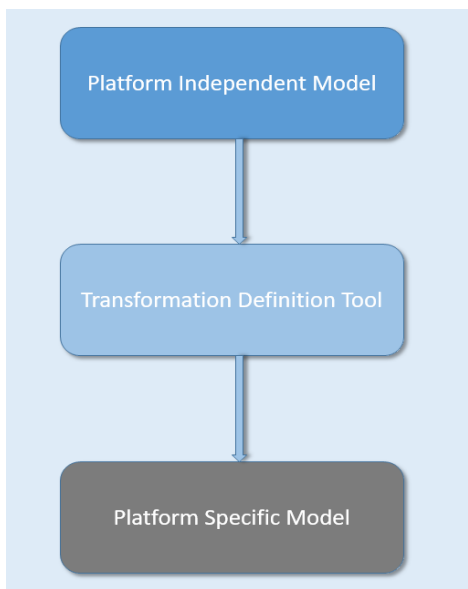


Fig. 2. Model Driven Architecture layers

MDA creates a new level of abstraction, above the implementation layers, in which the organizations are able to create their own objects to represent their business logic in an implementation independent way. The modelling language for this business abstraction layer is quite important for to identify the specific code part. The properties which describes the requirements are vital to identify the similar business processes and attach them to the implemented code parts.

B. Platform Independent Model definition

The implementations of a Platform Specific Model can be done in different ways, which is not the focus of this article. It focuses on finding an optimal way how the Platform Independent Model can be used to decide the common set of objects between the upgrade paths.

From the MDA point of view, three abstraction layers were successfully identified for SaaS implementations. CIM could specify the requirements for the system in a high level. PIM can describe a platform independent manner of the operation, while hiding the platform technical operations. PSM defines the platform specific model and describes the low-level platform details. MDA is able to map a single PIM in the cloud to more different PSM, as the transformation goes from platform independent view to platform specific view of the solution.

UML (Unified Modeling Language) description was used mostly because of being a standard. CIM, PIM and PSM layers are described with the aid of UML. This paper will show the examples from Dynamics 365 cloud based version, and will compare it with the already existing on-demand version of the same ERP system.

CIM model definition must be extended with the right property definitions, which has to be matched to the requirements list. The requirements list of the business

processes has to be matched to the requirements model of the CIM without losing valuable details. Having the right requirements model in CIM is vital to go be able to identify, what kind of requirements has to satisfy the code representation in PSM, and how the code elements can be identified, which can stay untouched after the SaaS model transition? The following chapter will show an elementary example, where the PIM model could be built up, attach it to the PSM model via the necessary transformation tool.

When analyzing the current codebase in the on premise version, part of the UML class schema looks like this:

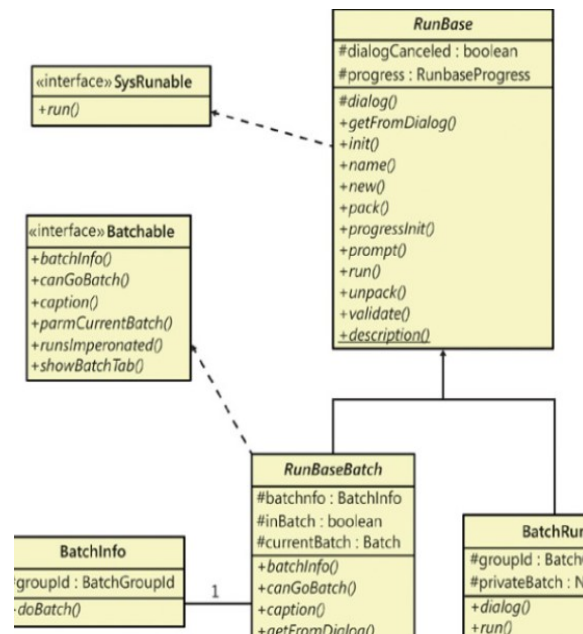


Fig. 3. Class schema of the RunBase class on Premise PIM model [6]

This UML model contains entity relationship data model, with all the elements of shared and private objects. It also contains code classes, interfaces, with all the attributes, operations, parameters types and return types. This PIM model is enough to decide the question: how can the two-object model compared during the upgrade process, whether they are compatible?

During the code upgrade process, only the currently existing business processed are in focus of refactoring, new business requirements are certainly not. New features, which can replace old features, are also not affected. Refactoring only affect the common set of business requirements, so that will also decrease the number of possible objects.

When the object hierarchy diagram is ready in UML format, the specific object is described by state of the object described by the attributes, and the behavior of the object described by the operations. These object are working in association and dependency, maybe interacting with other objects in their dependency graph. Aggregation and composition means a stronger level of association, with objects owning parts from other objects.

Composition is a higher level relationship than aggregation, meaning the whole part relationship.

III. BUSINESS PROCESS UPGRADE USING PIM MODEL BASED APPROACH

In order to show the possibility of reducing the necessary code changes between code upgrade from on premise version of cloud based SaaS version of the same software solution, the previously mentioned example will be examined.

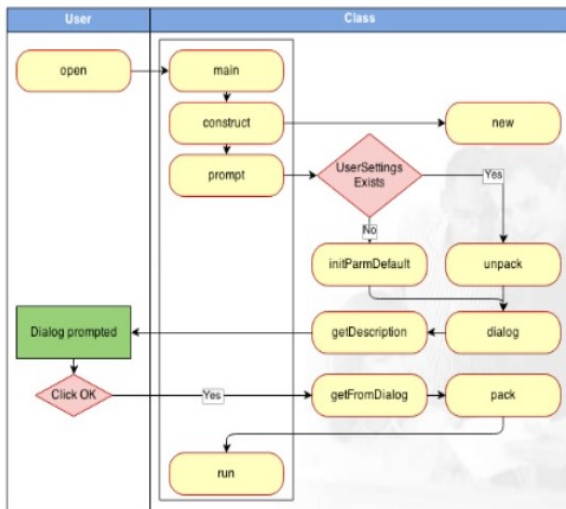


Fig.4. Runbase class operations in on premise PIM model [10]

The PIM for a SaaS cloud software solution in UML schema format contains objects, their attributes and operations. Accessor methods of the attributes are not explicitly described, because this task will be done during the MDA mapping, where this attributes will be translated into the right accessor methods. The PIM model will give a static view of the system.

The other main part, which can help identify the code parts responsible for implementing the similar requirement is the attributes: what are the input/return data formats, how the code interacts with its environment [17]?

All of these business requirements remain the same when upgrading from on premise version to SaaS cloud based version, so the first supposition is fulfilled. The requirements are the same, which means no new business logic has to be implemented, it is possible that no code change is needed by the PIM model. According to the PIM model, that the parameters and the operations needs to be the same in both software versions

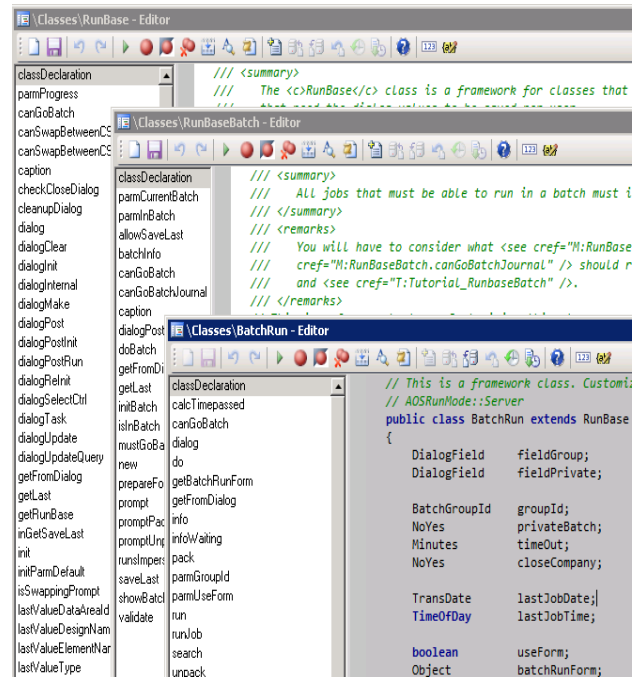


Fig 5. PSM implementation of the RunBase class

The PSM implementation of the RunBase and its connected classes are frequently used in running scheduled batch jobs. Main business requirements are the following:

- Schedule a batch job per user, company, batch group. Scheduling can be repetitious also.
- User interaction dialog for setting up the initial parameters
- Result messages logging
- Output actions handling, like format conversion, emailing, etc.
- Exception handling and error logging for late usage
- Restricted resource using through batch groups
- Service-like running
- Parallelism

Platform Specific Model can be implemented in different ways, this article does not bother with different implementation strategies. The focus of the article is finding a way where the Platform Independent Model is used to decide the common set of objects between the upgrade paths.

As for the comparison between the two PIM models, they seem similar. Generally, it goes by the expectations, because the business requirements are the same, so the expected outcome is to have similar operations model. The exact comparison for a couple of thousands of objects has to be done automatically.

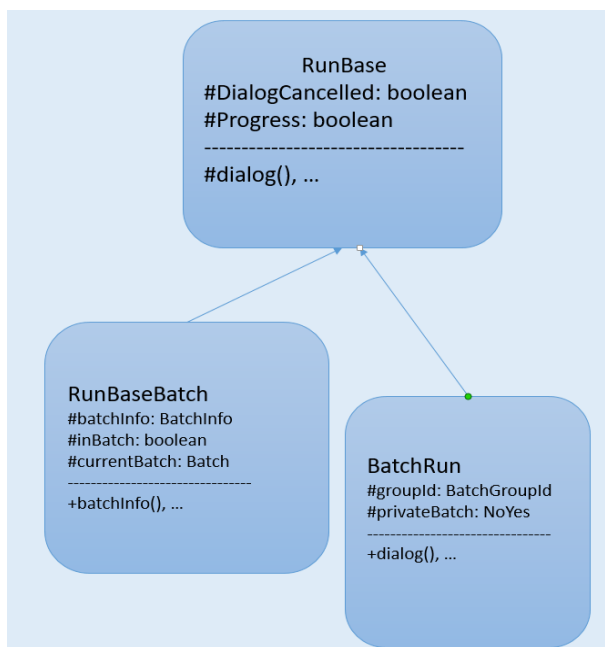


Fig.6. Runbase class operations in cloud base SaaS PIM model

This example shows, that PIM model stores enough information to decide about the refactoring of objects. As the Platform Specific Implementation layer is not touched, the implementation can be done in various ways, the process and the used models in PIM will not reflect any system specific properties of the software solution.

As the SaaS model brought new abstraction layer [16] into the cloud based software operation model, the refactoring becomes more vital, as the core business logic is the same or very similar. This new abstraction layer divides the core business logic from any transformation on infrastructure specific layers, and allows the customer focusing only with its business processes.

Although the underlying infrastructure seems to have longer lifecycle, with the aid of the new SaaS operating model, the software product can “live longer”. The basic business processes, like incoming invoice approval and posing a transaction into the General Ledger seems more or less stable in the past decades, well before the introduction of the Enterprise Resource Planning Systems [14]. This SaaS operating model makes it possible for the customers to always use the latest version of the software, without taking attention of the underlying infrastructure, which covers the following: applications runtime, middleware, operating system, virtualization, servers, storage, networking, security [15]. Not only the workload and the responsibility goes to anywhere else, but the customers has the possibility to deal with the only thing they can do better: tailoring the core business logic.

VI. CONCLUSIONS AND FUTURE WORK

PIM model is enough to decide about the code refactoring, if the object hierarchy is well defined, and the requirements list is up to date according to the business

requirements[11][12]. It contains every information, about the properties and the behavior of the code objects, so on the decision side, no need to bother with the transformation and underlying layers. The upgrade task from on premise to SaaS cloud based versions used the same middleware, OS and virtualization technology, so the PIM->PSM transformation was easy, because the same software technology.

Future work will be go towards to analyze how the different PSM models and the transformation affects the implementation of the same PIM model. How can be assured, that the transformation to different PSMs do not changes the core business logic implementation?

REFERENCES

- [1] Microsoft Dynamics Sure Step Methodology, <https://technet.microsoft.com/en-us/library/aa496439.aspx>
- [2] Richard Murch, The Software Development Lifecycle - A Complete Guide, Amazon Digital Services LLC, ASIN: B007ZCRP1I.
- [3] Clements, Kazman, Klein, Evaluating Software Architectures: Methods and Case Studies, 2002, ISBN-10: 020170482X
- [4] Alp Oral, Bedir Tekinerdogan, "Supporting Performance Isolation in Software as a Service Systems with Rich Clients", Big Data (BigData Congress) 2015 IEEE International Congress on, pp. 297-304, 2015.
- [5] Armando Fox, David A. Patterson: Engineering Software as a Service: An Agile Approach Using Cloud Computing, Strawberry Canyon LLC, 2013, ISBN 0984881247, 9780984881246
- [6] <https://www.microsoftpressstore.com/articles/article.aspx?p=2240847&seqNum=10>, The MorphX Development Environment and Tools
- [7] Peter Mell and Tim Grance: The NIST Definition of Cloud Computing, Version 15, 10-7-09, <http://thecloudtutorial.com/nistcloudcomputingdefinition.html>
- [8] Enterprise Resource Planning software blog, <http://www.erpsoftwareblog.com/2016/09/hosting-microsoft-dynamics-cloud-evaluating-iaas-paas-saas/>
- [9] Database Entity Relationship Diagramm, <https://community.dynamics.com/ax/b/axhari/archive/2014/08/09/ax-2012-database-entity-relationship-diagrams>
- [10] <https://www.slideshare.net/HamdaouiAmine/microsoft-dynamics-ax2012-forms-and-tables-methods-call-sequences-30159669>
- [11] Aggarwal, K. K., et al. "Software reuse metrics for object-oriented systems." Software Engineering Research, Management and Applications, 2005. Third ACIS International Conference on. IEEE, 2005.
- [12] Jeffrey S. Poulin, Ph.D., "The Search for a General Reusability Metric", Lockheed Martin Federal Systems Owego, New York, Proceedings of the Workshop on Reuse and the NASA Software Strategic Plan, Fairfax, VA, 24-27 September 1996.
- [13] Martin L. Griss: Systematic Software Reuse: Architecture, Process and Organization are Crucial, Software Technology Laboratory, HP Laboratories, Fusion Newsletter, <http://martin.griss.com/pubs/fusion1.htm>
- [14] István Orosz, László Szívós: The role of data authentication and security in the audit of financial statements, Acta Polytechnica Hungarica, ISSN 1785-8860, 2013
- [15] Krešimir Popović, Željko Hocenski: "Cloud computing security issues and challenges", MIPRO 2010 Proceedings of the 33rd International Convention, ISBN: 978-9-5323-3050-2
- [16] István Orosz, Tamás Orosz: Code reusability in cloud based ERP solutions, International Symposium on Intelligent Systems and Informatics (SISY 2017) ISBN: 978-1-5386-3854-5
- [17] A.Selmeci, I. Orosz, T. Orosz: Workflow processing using SAP Objects, Acta Cybernetica ISSN 0324-721X, 2015

Semantics-Preserving Encryption for Computer Networking Related Data Types

Gergő Ládi

Laboratory of Cryptography and System Security
 Department of Networked Systems and Services
 Budapest University of Technology and Economics
 Budapest, Hungary
 me@gergoladi.me

Abstract — Semantics-preserving encryption methods are encryption methods that not only preserve the format (data structure) of the input, but also a set of additional properties that are desired to be preserved (for example, transforming an IP address into another from the same subnet). Such methods may be used to anonymize logs or otherwise hide potentially sensitive information from third parties, while preserving characteristics that are essential for a given purpose. This paper presents tuneable semantics-preserving encryption methods that may be applied to common computer networking related data types such as IPv4, IPv6, and MAC addresses.

Keywords – semantics-preserving encryption; format-preserving encryption; networking; data type; MAC address; IPv4 address; IPv6 address; TCP port; UDP port; privacy; log anonymization;

I. INTRODUCTION

Format-preserving encryption methods, i.e. methods that – using a secret key – reversibly map an element of a set to another element of the same set, are most often used when data is to be shared with untrusted third parties that perform format checks on the input. In these cases, the output of a traditional encryption function would not be accepted by the third party, as the format checks would fail. In recent years, format-preserving encryption methods have been proposed for simple data types, such as integers [1] or *DateTimes* [2], as well as more complex ones, such as JPEG or PNG images ([3], [4]).

Computer networking related data types (e.g. IPv4 and IPv6 addresses) are usually treated as binary data, which works fine as long as only basic (length) checks are in place. However, several addresses and address types have special or additional meanings, therefore, an address encrypted into one of these may fail more sophisticated checks. This is where semantics-preserving encryption methods enter the picture.

Using semantics-preserving encryption, it can be ensured that alongside encrypting in a format-preserving manner, a set of desired properties are also transferred to the ciphertext. This property set should be tailored to each use case to contain the least amount of information to make validation checks pass, without revealing unnecessarily much. In addition, semantics-preserving encryption can also be used to anonymize logs containing computer network related data types. Then, these logs may be shared with third parties, without having to worry about leaking sensitive information such as the network

topology or internal addressing practices. Since this is encryption, the data owner may reverse the encryption using the key if needed.

This paper enumerates the most frequently used computer networking related data types, providing a semantics-preserving encryption method for each one.

II. RELATED WORKS

There are currently no known papers that address the topic of encrypting MAC addresses in a semantics-preserving manner. For IPv4 addresses, there exists *Tcpdpriv* [5], an open source anonymizer that may be configured to preserve the first N bits of the address, but it does not consider special addresses and address spaces. Xu et al. [6] show that it also has the drawbacks of not being consistent between different traces, and that it has a rather large memory footprint (320 MBs of memory for a trace with 10 million unique IPv4 addresses). *Tcpdpriv* did not originally support IPv6 addresses, but was extended by Cho et al. [7] to do so. There exist additional proposals ([8], [9], [10]) for transforming IPv4/IPv6 addresses, but none of them goes further than preserving prefixes, and some of these methods are not even reversible.

III. MATHEMATICAL BACKGROUND

This section explains the mathematical background that is needed to understand how format-preserving encryption, the basis of semantics-preserving encryption works.

All of the proposed algorithms herein rely on the existence of a function \mathcal{F} , where \mathcal{F} is a format-preserving transformation (encryption) as defined by Black and Rogaway [11], where \mathcal{K} is the encryption key, and \mathcal{M} is the message space:

$$\mathcal{F}: \mathcal{K} \times \mathcal{M} \rightarrow \mathcal{M} \quad (1)$$

For \mathcal{F} , an inverse function \mathcal{F}^{-1} must also exist such that if (2) holds true, then (3) is also true:

$$\forall k \in \mathcal{K}; \forall m_1, m_2 \in \mathcal{M}; \mathcal{F}(k, m_1) = m_2 \quad (2)$$

$$\mathcal{F}^{-1}(k, m_2) = m_1 \quad (3)$$

In their paper, *Format Preserving Encryption*, Bellare et al. [12] construct functions that work in this manner, but \mathcal{M} can only be a set of integers of arbitrary width (often referred to as

BigIntegers). These functions can be used as the basis of a rank-then-encipher approach that may be used to construct others that work on sets of arbitrary data types. Rank-then-encipher methods have three distinct steps:

- Ranking: consider a set \mathcal{M} , where \mathcal{M} can now contain elements of arbitrary data types (not just integers). Define a bijective mapping G (eq. 4):

$$G: \mathcal{M} \rightarrow \{0, \dots, |\mathcal{M}|-1\} \subset \mathbb{N} \quad (4)$$

Apply G to the value m to be encrypted:

$$rank = G(m \in \mathcal{M}) \quad (5)$$

Note: In practice, G can be a simple function that maps each element of \mathcal{M} to its index (position) in the set (i.e. the first element maps to 0, the second one to 1, and the last one to $|\mathcal{M}|-1$).

- Encipherment: use one of Bellare's functions (with any key of our choice) to encrypt the rank from the previous step:

$$enc_index = FE2_Enc(k \in \mathcal{K}, rank, |\mathcal{M}|-1) \quad (6)$$

- Unranking: since G is bijective, it has an inverse G^{-1} . Applying G^{-1} to enc_index yields an element of \mathcal{M} , thus we have constructed an $\mathcal{M} \rightarrow \mathcal{M}$ function.

Note: If the above-mentioned simple function was used for ranking, then G^{-1} essentially just returns the enc_index^{th} element of \mathcal{M} .

Decryption works similarly:

- Ranking: apply G to the value c to be decrypted, just as in eq. (5).
- Decipherment: use the inverse of Bellare's functions with the same key k as in eq. (6) to decrypt the $rank$ from the previous step:

$$dec_index = FE2_Dec(k, rank, |\mathcal{M}|-1) \quad (7)$$

- Unranking: applying G^{-1} to dec_index returns the original m .

Making use of the rank-then-encipher method, it is now possible to construct algorithms that in addition to preserving the format of the input, they also preserve semantics.

For the rest of this paper, unless otherwise mentioned, encryption will refer to encrypting in a format-preserving manner, while decryption will refer to reversing said encryption.

IV. PRESERVING SEMANTICS FOR COMPUTER NETWORKING RELATED DATA TYPES

In order to design semantics-preserving methods, it is imperative to understand how each kind of address is structured and which addresses or address ranges have special meanings (this is what needs to be preserved). This section aims to describe the most commonly used networking-related

data types, then proposes algorithms that may be used to encrypt said data types in a semantics-preserving manner.

A. MAC Addresses

Media Access Control (MAC) addresses are 48-bit physical addresses that are mostly used by network devices running IEEE 802-based network technologies, such as Ethernet, WiFi, or Bluetooth. The 6-byte address comprises of two 3-byte parts (see Figure 1.): the Organizationally Unique Identifier (OUI), and the Network Interface Controller specific identifier [13].

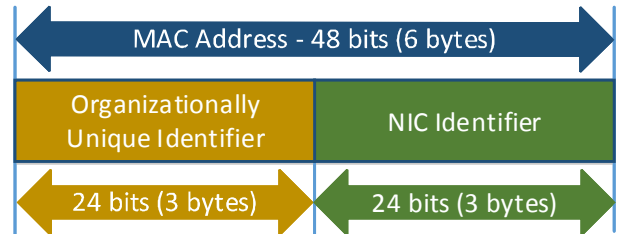


Figure 1. Structure of a MAC address

Generally, OUIs are allocated by the IEEE Registration Authority, and are used to uniquely identify the company that manufactured the network interface card, while NIC identifiers are assigned by the manufacturers under the requirement that every single NIC ever produced should be assigned a unique identifier [13]. The first and second least significant bits of the most significant byte of the OUI have special meanings (Figure 2.): the least significant bit (Individual/Group bit) denotes whether the MAC address is a unicast (value of 0) or multicast/broadcast (value of 1) address, and the second least significant bit (Global/Local bit) denotes whether the address is locally administered (overridden by the system administrator) (value of 1) or not (value of 0).

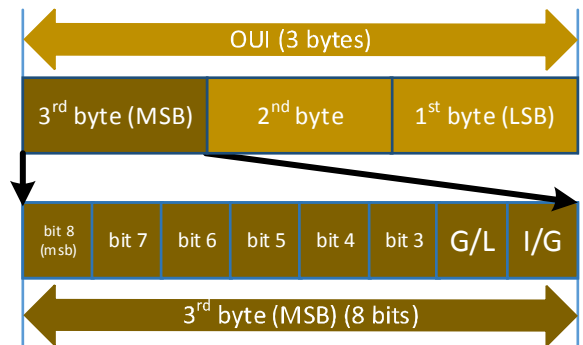


Figure 2. Position of the G/L and I/G bits in the OUI

To preserve semantics, the value of these two special bits should be preserved. For locally administered addresses, the remaining 46 bits of the address may be encrypted without restrictions, whereas for unique addresses, two different strategies may be used:

- Leave the OUI as is, and only encrypt the NIC identifier. This ensures that the encrypted address will appear to be one from the same vendor as the original.

TABLE I. SPECIAL-USE MAC ADDRESS BLOCKS

Block start	Block end	Defined in	Purpose
00:00:5E:00:00:00	00:00:5E:00:00:FF	RFC 7042	Reserved, ratification required for assignment
00:00:5E:00:01:00	00:00:5E:00:01:FF	RFC 5798	Virtual Router Redundancy Protocol (VRRP) for IPv4 networks
00:00:5E:00:02:00	00:00:5E:00:02:FF	RFC 5798	Virtual Router Redundancy Protocol (VRRP) for IPv6 networks
00:00:5E:00:52:00	00:00:5E:00:52:FF	RFC 7042	Reserved for small allocations (3 addresses currently allocated [14])
00:00:5E:00:53:00	00:00:5E:00:53:FF	RFC 7042	For use in documentation only (unicast)
01:00:5E:00:00:00	01:00:5E:7F:FF:FF	RFC 1112	IPv4 multicasting
01:00:5E:80:00:00	01:00:5E:8F:FF:FF	RFC 5332	Multi-Protocol Label Switching (multicast)
01:00:5E:90:00:00	01:00:5E:90:00:FF	RFC 7042	Reserved for small allocations (4 addresses currently allocated [14])
01:00:5E:90:10:00	01:00:5E:90:10:FF	RFC 7042	For use in documentation only (multicast)
01:80:C2:00:00:00	01:80:C2:FF:FF:FF	802.1D	Reserved for use in the 802.1D IEEE standard for MAC bridges
33:33:00:00:00:00	33:33:FF:FF:FF:FF	RFC 2464	IPv6 multicasting

- Make a list of valid OUIs¹, then use a rank-then-encipher function to encrypt the OUI. Encrypt the NIC identifier separately. This ensures that the encrypted address will appear to be from a valid manufacturer.

Some address blocks, instead of being assigned to vendors, are reserved for use in specific protocols (see Table I.). In order to encrypt addresses from any of the reserved blocks, it is important to understand how those addresses are allocated and used. For the VRRP MAC ranges, the last octet comes from the VRID², which also appears in upper-layer packets, therefore care should be taken to encrypt it consistently everywhere where it appears. Small allocation ranges at present have only single address allocations with a specific purpose [14], these should not be changed during encryption. The ranges reserved for documentation should never be used in a non-hypothetical network, therefore these need not be encrypted (but if desired, the last octet may be encrypted freely). For the MAC range used by MPLS, the last 20 bits of the address are calculated from MPLS labels³, meaning that these should only be encrypted along with the labels to preserve consistency. The address range reserved for 802.1D consists of single allocations with a specific purpose [15], these addresses should be left intact. As for the IPv4 and IPv6 multicast MAC ranges, the lowest 23 and 32 bits (respectively) are derived from the IP addresses, therefore these should not be directly encrypted, but instead, be recalculated when the IP addresses are encrypted.

There are also two MAC addresses with a special meaning: the so-called *unspecified* address 00:00:00:00:00:00 (with all the bits set to 0), and the broadcast address ff:ff:ff:ff:ff:ff (with all the bits set to 1). These should not be changed during encryption.

¹ The list changes frequently, and is too long to be included here. The latest version may be acquired from <https://standards-oui.ieee.org/oui/oui.txt>.

² Virtual Router Identifier, a number used by the VRRP protocol.

³ A 20-bit field used for routing packets in MPLS networks.

Considering the previous observations, the following algorithm can be given to encrypt MAC addresses in a semantics-preserving manner:

- 1) If the input is the unspecified address or the broadcast address, return the input unchanged.
- 2) If the input is in one of the special-use blocks, encrypt the address (and possibly other fields in the packet, as needed) based on the previous considerations.
- 3) If the address has the locally administered bit set, consider the address without the I/G and G/L bits as a 46-bit integer, then encrypt it. In the resulting ciphertext, re-insert the original I/G and G/L bits in the appropriate positions. Return this 48-bit number as a MAC address.
- 4) Otherwise, use one of the OUI encryption strategies mentioned earlier, then encrypt the NIC identifier. Construct a new MAC address from these two ciphertexts.

The decryption process can also be described in the same fashion:

- 1) If the input is the unspecified address or the broadcast address, return the input unchanged.
- 2) If the input is in one of the special-use blocks, decrypt the address and any other fields that were changed during encryption.
- 3) If the locally administered bit is set, consider the address without the I/G and G/L bits as a 46-bit integer, then decrypt it. In the resulting plaintext, re-insert the original I/G and G/L bits in the appropriate position. Return this 48-bit number as a MAC address.
- 4) Otherwise, we have to know which strategy was used for encrypting the OUI. If the OUI was

TABLE II. SPECIAL-USE IPV4 ADDRESS BLOCKS

Block	Defined in	Purpose	Block	Defined in	Purpose
0.0.0.0/8	RFC 1122	<i>Unspecified</i> address	192.0.2.0/24	RFC 5737	Documentation (TEST-NET-1)
10.0.0.0/8	RFC 1918	Private use	192.168.0.0/16	RFC 1918	Private use
100.64.0.0/10	RFC 6598	Carrier-grade NAT	198.18.0.0/15	RFC 2544	Benchmarking
127.0.0.0/8	RFC 1122	Loopback addresses	198.51.100.0/24	RFC 5737	Documentation (TEST-NET-2)
169.254.0.0/16	RFC 3927	Link-local addresses	203.0.113.0/24	RFC 5737	Documentation (TEST-NET-3)
172.16.0.0/12	RFC 1918	Private use	240.0.0.0/4	RFC 1112	Reserved
192.0.0.0/24	RFC 6890	IETF assignments	255.255.255.255/32	RFC 1122	Limited broadcast

preserved, keep the OUI and decrypt the NIC identifier. If the OUI was not preserved, reverse the rank-then-encipher method on the OUI over the same list that was used during encryption, then decrypt the NIC identifier.

B. IPv4 Addresses

Internet Protocol version 4 (IPv4) addresses are 32-bit logical addresses that are used by devices to communicate over IPv4 networks. In order to encrypt IP addresses in a semantics-preserving manner, it should be understood how IP addresses are allocated, formed, and used. The addresses are assigned in blocks by the Internet Assigned Numbers Authority (IANA), either to be leased to customers through other organizations, or to be used for special purposes [16].

IPv4 addresses are usually written in the form A.B.C.D, where A, B, C, and D are integers between 0 and 255. Each address is made up of network bits and host bits – the former identify the network, the latter a host. The number of network bits may be given in the Classless Inter-Domain Routing (CIDR) notation, where it is appended to the IP address following a forward slash (e.g. 127.0.0.1/8), or it may be given in the form of a subnet mask (e.g. 255.0.0.0), where each 1-bit in the mask means that the corresponding bit in the IP address identifies the network. The first and last addresses of each subnet are special: the first one is the *network address*, while the last one is the *broadcast address* for that subnet – these may not be assigned to hosts. Networks usually have a *default gateway* – a host to which other hosts in the same network can send packets that are addressed to hosts on different networks. While there is no standard for this, default gateways are usually assigned the first or last assignable address of the subnet.

During encryption, care should be taken never to transform an IP address that is not a network address into one that is, to transform an address that is not a broadcast address into one that is, and vice versa. If the size of the network is known and is relevant (e.g. when anonymizing an internal network), it should be ensured that if any two hosts were in the same subnet before encryption, they should remain in the same subnet after encryption (but the subnet itself, of course, may be different). If there was a default gateway, and it had the first or the last address, it should have the first or last address in the anonymized (encrypted) network as well.

When encrypting IP addresses allocated for special purposes (see Table II.), to preserve semantics, it should be understood how the addresses in those blocks are used. Of the 0.0.0.0/8 block, usually only the address 0.0.0.0 is used – this should not be changed during encryption, while the rest of the addresses in the block should encrypt into other addresses in the block. The same applies for the loopback range – usually only the first address, 127.0.0.1 is used, this should not be transformed, but the rest of the addresses in the block should be. Addresses in the 100.64.0.0/10 block should be transformed into addresses in the same block. Private use blocks are freely assignable by network administrators, these addresses are used in internal networks and do not appear on the internet. Subnets in these blocks should only be transformed into other subnets in these blocks, but the subnets do not have to remain in the same private block (for example, transforming 192.168.10.0/24 into 172.16.20.0/24 is allowed). Addresses in the link-local block should be transformed into addresses in the same block, except for the first and last 256 addresses, which are reserved for later allocation [17], and should be left as is. Addresses in the IETF-reserved block should not be transformed. Addresses in the documentation blocks should never be used in actual networks. These may be transformed into addresses from the same block or may be left as is; in addition, each block as a whole may be transformed into one of the other blocks. Addresses in the benchmarking block may be left as is, or may be transformed into addresses from the same block. Addresses from the 240.0.0.0/4 (reserved) block should never appear anywhere, deciding how to deal with these is left for the implementor of the encryption algorithm. Finally, the limited broadcast address, 255.255.255.255 should be left intact.

There also exists a block, 224.0.0.0/4 that is used for point-to-multipoint (multicast) traffic [18]. The block contains several single-address allocations, range allocations, as well as ranges of reserved and unassigned addresses. These are not listed in this paper due to lack of space and frequency of updates to the allocations. When encrypting addresses from this block, single-address allocations should not be transformed, while addresses from range allocations should be encrypted into other addresses from the same range. Reserved addresses and address ranges should not appear in logs, but may be treated as single and range allocations, respectively. Unassigned addresses and ranges should be treated as single and range allocations, respectively.

TABLE III. SPECIAL-USE IPV6 ADDRESS BLOCKS

Block	Defined in	Purpose	Block	Defined in	Purpose
::/128	RFC 4291	<i>Unspecified</i> address	2001::/32	RFC 4380	TEREDO (4-to-6 transition technology)
::1/128	RFC 4291	Loopback address	2001:2::/48	RFC 5180	Benchmarking
::ffff:0:0/96	RFC 4291	IPv4-mapped addresses	2001:20::/28	RFC 7343	ORCHIDv2
64:ff9b::/96	RFC 6052	IPv4-IPv6 translation	2001:db8::/32	RFC 3849	Documentation
64:ff9b:1::/48	RFC 8215	IPv4-IPv6 translation	2002::/16	RFC 3056	6to4 (transition technology)
100::/64	RFC 6666	Discard addresses	fc00::/7	RFC 4193	Private use (unique local)
2001::/23	RFC 2928	IETF assignments	fe80::/10	RFC 4291	Link-local unicast

Transforming addresses into other addresses in the same block is done by leaving the network bits as is, then using a rank-then-encipher function on the host bits (section III.), modified to exclude the network and broadcast addresses from the mapping. Where addresses have to be transformed among a list of blocks, a rank-then-encipher function may be used on the list index for the network bits, and the host bits can be transformed as explained previously.

Considering all of the above, encrypting an IPv4 address in a semantics-preserving manner can be done using the following steps:

- 1) If the address must be left intact, leave it as is.
- 2) If the address is in one of the special-use blocks, encrypt it based on the previous considerations.
- 3) If the number of network bits is known, leave or encrypt the network bits as needed, then encrypt the host bits as explained previously.
- 4) Otherwise, the address is from a public, non-special range and nothing else is known about it, so make a mapping of addresses in non-special ranges, then use a rank-then-encipher function to encrypt the address.

Decryption is as follows:

- 1) If the address was to be left intact, do nothing.
- 2) If the address is in one of the special-use blocks, decrypt it based on the previous considerations.
- 3) If the number of network bits is known, decrypt the network bits if they were encrypted previously, then decrypt the host bits.
- 4) Otherwise, the address was from a public, non-special range and nothing was known about it, so use the same mapping as what was used for encryption, then use the rank-then-encipher function to decrypt the address.

C. IPv6 Addresses

Internet Protocol version 6 (IPv4) addresses are 128-bit logical addresses that are used by network devices to

communicate over IPv6 networks. Addresses are written in the form of A:B:C:D:E:F:G:H/n, where A to H are 16-bit blocks represented as hexadecimal values (e.g. fe80) and *n* is the prefix length, having the same function as subnet masks in IPv4. Since the addressing works similarly to IPv4 addresses, the algorithms for encrypting and decrypting these addresses semantics-preservingly are also similar, with a few notable exceptions.

In IPv6, there is no broadcasting. As a result, there are also no broadcast addresses. This means that the last address of a prefix no longer has to be treated in a special way.

As with IPv4, some address blocks (see Table III.) are reserved for special use [19], these need to be treated differently. The unspecified and the loopback address ranges have been reduced to one address each, these should not be changed during encryption. The documentation, the benchmarking, the IETF-assigned range, and the private block should be treated as they were in IPv4. Addresses from the discard block and the link-local unicast block should be transformed into other addresses from the same block. ORCHIDv2 addresses are only to be used as non-routable overlay addresses, thus should not appear in regular traffic logs – in case they do, they need to be changed into a valid ORCHID identifier (this is not discussed further in this paper).

Addresses in the IPv4-mapped address block, the TEREDO block, the IPv4-IPv6 translation blocks, and the 6to4 block contain IPv4 addresses, either embedded in the IPv6 address, or in the payload. When encrypting addresses from these blocks, transform the embedded address instead of the IPv6 address.

In addition, IPv6 has a multicast address block, ff00::/8 [20]. This should be treated in the same way as addresses in the IPv4 multicast block.

Finally, some IPv6 addresses are generated from the interfaces' MAC address using the EUI-64 algorithm [21]. For cases when both the MAC and the IPv6 addresses are to be encrypted and the IPv6 address appears to have been calculated from the MAC address, encrypt the MAC address first, then recalculate the IPv6 address using the resulting ciphertext.

D. TCP and UDP Port Numbers

Port numbers are 2-byte numeric identifiers that are used to identify network services running on hosts. Port numbers belong to one of the three categories [22]:

- 1) Well-known ports (from 0 to 1023),
- 2) Registered ports (from 1024 to 49151),
- 3) Dynamic (also referred to as private) ports (from 49152 to 65535)

The list of well-known and registered ports is maintained by IANA. Ports in these ranges generally indicate a specific service running on a host (e.g. TCP port 25 identifies an SMTP service that can be used to send mail). However, this list is a recommendation, not a requirement, and any service may be set to listen on any port. It is also worth noting that some ports in these ranges are unassigned or reserved.

To encrypt port numbers, the following algorithm can be used:

- 1) If the port is a dynamic port, encrypt it into another from the dynamic range.
- 2) If it is a well-known or registered port that is not unassigned or unallocated, leave it as is.
- 3) Otherwise, make a list of unassigned and unregistered ports, then use the rank-then-encipher approach to encrypt it.

The process of decryption is not described due to the lack of space, but it may be easily inferred by looking at the algorithm used for encryption.

V. FURTHER CONSIDERATIONS

When relying on the lists of reserved OUI blocks for MAC addresses and reserved IP blocks for IPv4/IPv6 addresses, one should check the IANA allocation tables ([14], [16], [18]-[20]) to ensure that the latest, most up-to-date list is used.

CONCLUSION

Semantics-preserving encryption is useful for a variety of use cases where data has to be encrypted or anonymized before being shared with third parties while still preserving certain properties of the source. Despite having practical uses, no elaborate solutions have been offered to date. The aim of this work was to demonstrate what semantical information the various computer networking related data types carry, then present algorithms that can be used to encrypt these data types while preserving this semantical information.

REFERENCES

- [1] M. Bellare, T. Ristenpart, P. Rogaway, T. Stegers, "Format-Preserving Encryption," Cryptology ePrint Archive: Report 2009/251 (online: <https://eprint.iacr.org/2009/251>), 2009
- [2] Z. Liu, C. Jia, J. Li, X. Cheng, "Format-preserving encryption for DateTime," IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2010
- [3] A. Unterweger, A. Uhl, "Length-preserving bit-stream-based JPEG encryption," MM&Sec '12 Proceedings of the 14th ACM multimedia and security workshop, 2012, pp. 85-90.
- [4] Z. Liu, M. Li, X.-Y. You, C. Jia, "Format-preserving encryption for PNG image," Beijing Ligong Daxue Xuebao/Transaction of Beijing Institute of Technology, 2013
- [5] G. Minshall, Tcpdpriv, <http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html> (online), 1996
- [6] J. Xu, J. Fan, M. Ammar, S. B. Moon, "On the Design and Performance of Prefix-Preserving IP Traffic Trace Anonymization", IMW '01 Proceedings of the 1st ACM SIGCOMM Workshop on Internet measurement, 2001, pp. 263-266
- [7] K. Cho, K. Mitsuya, A. Kato, "Traffic Data Repository at the Wide Project," Proceedings of the FREENIX Track: 2000 USENIX Annual Technical Conference, 2000
- [8] A. J. Slagell, Y. Li, K. Luo, "Sharing Network Logs for Computer Forensics: A New Tool for the Anonymization of NetFlow Records," Workshop of the 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks, 2005
- [9] D. Plonka, A. Berger, "kIP: a Measured Approach to IPv6 Address Anonymization," Measurement and Analysis for Protocols Research Group (MAPRG) meeting, 2017
- [10] P. Zhang, X. Huang, M. Luo, C. Ning, Y. Ma, "Fast restorable prefix-preserving IP address anonymization for IPv4/IPv6," The Journal of China Universities of Posts and Telecommunications, 2010, pp. 93-98
- [11] J. Black, P. Rogaway, "Ciphers with arbitrary finite domains," RSA-CT, 2002, p. 114.
- [12] M. Bellare, T. Ristenpart, P. Rogaway, T. Stegers, "Format-Preserving Encryption," Cryptology ePrint Archive: Report 2009/251 (online: <https://eprint.iacr.org/2009/251>), 2009
- [13] IEEE Standards Association, "802-2014 – IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture," ISBN 978-0-7381-9219-2, pp. 22-27
- [14] Internet Assigned Numbers Authority, "Assignments, Ethernet Numbers," (online), <https://www.iana.org/assignments/ethernet-numbers/ethernet-numbers.xhtml>, retrieved: 2017/09/22.
- [15] IEEE Standards Association, "802.1D – IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges," ISBN 0-7381-3982-3 SS95213, p. 51.
- [16] Internet Assigned Numbers Authority, "IANA IPv4 Special-Purpose Address Registry," (online), <https://www.iana.org/assignments/iana-ipv4-special-registry/iana-ipv4-special-registry.xhtml>, retrieved: 2017/09/23.
- [17] S. Chesire, B. Aboba, E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses," (online), <https://tools.ietf.org/html/rfc3927>, Internet Engineering Task Force (IETF) RFC 3927, 2005, pp. 3, 24.
- [18] Internet Assigned Numbers Authority, "IPv4 Multicast Address Space Registry," (online), <https://www.iana.org/assignments/multicast-addresses/multicast-addresses.xhtml>, retrieved: 2017/09/23.
- [19] Internet Assigned Numbers Authority, "IANA IPv6 Special-Purpose Address Registry," (online), <https://www.iana.org/assignments/iana-ipv6-special-registry/iana-ipv6-special-registry.xhtml>, retrieved: 2017/09/24.
- [20] Internet Assigned Numbers Authority, "Internet Protocol Version 6 Address Space," (online), <https://www.iana.org/assignments/ipv6-address-space/ipv6-address-space.xhtml>, retrieved: 2017/09/24.
- [21] R. Hinden, S. Deering, "IP Version 6 Addressing Architecture," (online), <https://tools.ietf.org/html/rfc2373>, Internet Engineering Task Force (IETF) RFC 2373, 1998, p. 18.
- [22] Internet Assigned Numbers Authority, "Service Name and Transport Protocol Port Number Registry," (online), <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml>, retrieved: 2017/09/24.

Information Society supporting an optional chapter (the Wonderful World of Measurements) in secondary school physics

Cs. Fülöp

Madách Imre Gimnázium, Budapest, Hungary

Physics Education Program/ELTE PhD School of Physics, Budapest, Hungary

MTA-ELTE Physics Education Research Group, Budapest, Hungary

fulcsilla@gmail.com

It is needless to underpin that STEM (Science Technology Engineering Mathematics) is in crisis. What can we do in secondary education of science to ease the peril?

Today, instead of the teacher's presentation and interpretation, hands-on activities seem to appear on the palette of science methodology in secondary education. If a hands-on activity includes measurement too, our students face the tasks without adequate preparation. I am going to introduce a chapter specially worked out for secondary school students on measurement theory. I study what notions and methods are worthy to teach. Based on my practice I will also show motivating proposals for classroom use. I will focus on the use of the internet.

The chapter equips our students to study in RBL (Research Based learning) or SMP (Student's Measuring Project) methods, and to deeper understand what science is all about.

I. INTRODUCTION

The crisis of STEM is in the focus worldwide: in the US [1] the Department of Education, the Bureau of Labor Statistics and President Obama himself is urging a way out. The Hungarian model of science education in use about a hundred years ago, attracted the interest of the world. But studies show that the old glory has gone. We can say that it is like: the Hungarian model of science education is from triumph to trouble.

Representatives of active learning pedagogy, like R(earch) B(ased) L(earning) or S(tudents) M(easuring) P(roject) are worthy to pay attention to.

II. THE WONDERFUL WORLD OF MEASUREMENTS

A) *An optional unit for secondary education*

The history of science often is parallel of the individuals' progression in knowledge. It is mainly so in the field of scientific notions [2].

Galileo Galilei (1564-1642) is called "the father of science" as he recognised the need of validating the theory by experiments and measurements, and thus rethinking the key element in science, turning the focus of cognition on the scientific method. He set up standards for length and time. He made this to ensure that measurements done on different days or by different people can be compared in a reproducible way. We call this the Galilei-turn of scientific cognition.

Every course book prepared for secondary school education mentions the importance of measurements, usually in the introduction or in the first chapters; I could not find this content enough.

From my practice as a secondary school teacher I can conclude that despite some elements are usually in the curriculum, the first lessons are mainly dedicated to classroom management issues rather than making an overview of these notions.

The national curriculum gives a chance for the teacher to freely make the best use of 7 lessons (that is 10% of the course) each academic year on his discretion for the currently taught class. So, the unit I developed is adapted to this frame.

B) *The concept of measurement*

First, we need to define what measurement is. We can give a definition like this:

"Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events. The science of measurements is called metrology."

Measurement is an operation of a system, the so-called measuring system. It can be either simple or highly complicated. But the factors of it are these:

- object, something we want to get information about
 - device or meter, that is a measuring instrument
 - interaction between the object and the meter
- The system will provide the result of the measurement. This information is of three elements:
- a magnitude, which is a numerical value of the characterization
 - a unit, which is usually a standard, therefore the magnitude is the ratio of the measured quantity and this very standard
 - uncertainty, that inevitably comes from the operation of our measuring system.

In present days it is important to find and help our students to be good users of the information on the internet. A series of three videos are more than worthy to offer for students and teachers to learn about uncertainty: "Precision: measure of all things" [3]

C) Errors, not mistakes

I find that simple existence of uncertainty is hard to accept for secondary students. They find it easy to understand that the result is the duo of a magnitude and a unit.

When explaining the importance of the interaction as the crucial component of the operation of the system, uncertainty can come clear as an important factor in the result. Uncertainty indicates the confidence level of the measurement. Random and systematic errors are represented in it. We need to emphasize that errors are not mistakes.

Random error

If we measure a constant quantity several times we are likely to gain slightly different values. We call this type of error random, because it can't be predicted from the previous values. We should not use our instruments in their extremes of their operating limits, so we can reduce this error. Multiple measurements can help us to estimate this type of error: "making more measurements and calculating the average" gives us a more accurate result.

We can demonstrate simply this type of error by using a digital tool, like a kitchen scale. We can place a light or heavy (compared to its operating limit) object on it. The last digit varies; still, the students can see that we measure the very same object. We can perform 5 measurements. The reading is 19g (2 times) and 18g (3 times) for the same scone. It is in the lower extreme span of our tool, since it measures to 5000g.

Systematic error

The measured value contains an offset. It is an error that remains constant in a measurement setting. The measuring instrument determines the range of this type of error. The goal is to minimize it in the measuring procedure. We can use standardised protocols and instruments to meet our goal. It is also well-known as measurement or statistical bias. The sources of systematic error are: problematic calibration, the related manner of the measured quantity and drift.

We can also demonstrate this type of error with simple tools. The task is to measure the length of a rod with one provided tool. We can provide a meter rod, a ruler, a Vernier calliper or a micrometre screw. For our students, this point will be obvious.

Calibration and authentication

When discussing errors two notions arise. These are the process of authentication and calibration. Most of our students have a smart phone in their pockets. Using the internet, they can easily find out what these procedures are.

Authentication is an official analysis to prove that our device measures according to its protocol. Devices that are used in commerce must be subjects to the procedure regularly.

Calibration is a method to determine measuring characteristics of a given device. It a check to see if our measuring instrument is accurate.

Significant figures and error propagation

We can consider the accuracy of our data in secondary education by introducing the concept called

significant figures. The number of significant figures can be counted with a simple method: from the left we count the non-zero digits in the magnitude. Now, our students can understand that using devices of greater sensitivity means that the result of the measurement is more accurate, like the ones in applied scientific research [4].

To see if the students can understand the idea we can ask them to make a group discussion and based on it, to give a short reasoning (data should vary for the groups):

Your task is to interpret and show the difference in a practical example between these two results:

Peter: "The result is 3 decimetres."

David: "The result is 3.00 decimetres."

The answer is not a problem for our students. They come up with great examples rooting in their everyday life, like this reasoning:

"Both lads are gardening. Peter is a farmer; he is talking to his friends in a pub, explaining how his corps are growing. David is a student at Agricultural Secondary School; he is working on his project. He is talking to his teacher about how his corps are growing with the fertilizer he is studying."

If we work with significant figures, the rules of error propagation are also easy:

- 1) When we add (or subtract) we take the minimum of the significant figures.
- 2) When we multiply (or divide) we add the number of significant figures.

In the calculations, we use the rules of rounding. We can show how it goes on an example, Figure 1.

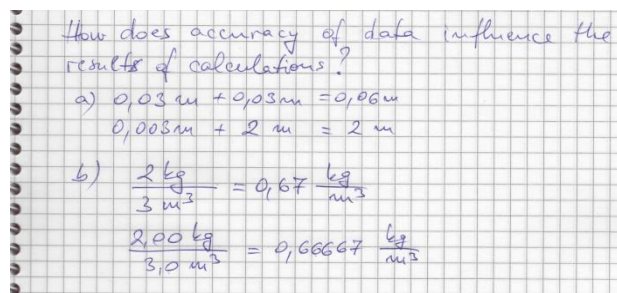


Figure 1.: Error propagation in public education

D) The International System of Units (SI)

First, I find it important to show that a need emerged for this system in history. In literature, we often meet the measure of cubic capacity. Therefore, our students have some units in their passive vocabulary from their reading experience. The first task I give for them is to look up what we mean by "akó" and "icce". They can use the internet to find the answer.

"akó": <https://hu.wikipedia.org/wiki/Ak%C3%B3>

"icce": <https://hu.wikipedia.org/wiki/Icce>

They can find that there were different measures of "akó". "Akó" in German is Eimer, meaning 58 litres until 1762, and 56.589 litres later. The Hungarian "akó" widens the picture as a great example to show how complicated the system of unit used to be:

-the Pest "akó" meant 53.72 litres until 1700, then 54.94 litres, and 58.6 litres

-the Buda "akó" meant 53.72 litres

The unit called “icce” is a suitable unit for further study: the Hungarian “icce” means 0.848 litres; the Pozsony “icce” means 0.839 litres, the Transylvanian “icce” 0.707 litres.

For students with a background in English culture I would give a similar task concerning “barrel” and “gallon”.

We can use old units for mass, area, volume, length, etc. Exciting items of information will enchant our students, and brings interest for them in topic. Some items I found most motivating for my students are:

- St. Steven, the first Hungarian king had a shoe-size of 48. We know it as the Hungarian foot is 31.26 cm (, whilst the English foot is 30.48 cm).
- In the Hungarian Great Plain, a unit of the area was “sheep”, referring to the area a sheep can graze in one day.
- The unit “mile” is the distance one can walk without a halt. Depending on the individual’s fitness it means different distances. The variations of this unit worldwide range from 1.7 to 11.3 kilometres.

The need of a consequent and coherent system for measures and units is obvious.

The early history of SI dates the French Revolution (the 1790s).

The basic requirements of the system were laid in 1860s in the British Association for the Advancement of Science: the system should contain base units and derived units. Base units are to be taken from nature. There are 7 mutually independent quantities and their units in SI. All other units are to be derivatives from these. The base units are materialized in the forms of standards. A set of standards were made and one of those were selected by random to be appointed as the international standard or prototype. These are kept in Sevres, in France. The rest of the standards were raffled among the nations that joined this system, they are the national standards. The Hungarian standards are in the OMH, an institution we have talked about already. Interestingly, the only western nation that has not adopted this system is the USA.

TABLE I.
SI BASE QUANTITIES ADJUSTED TO PUBLIC EDUCATION

quantity	symbol	unit	symbol
length	l	metre	m
mass	m	kilogram	kg
time	t	second	s
electric current	I	amp	A
charge	Q	coulomb	C
temperature	T	Kelvin degree Celsius	K °C
amount of substance	n	mole	mol
amount	N	piece/bit/---	--
luminous intensity	I	candela	cd

The seven base quantities and their units are shown in Table I. I inserted some quantities in green to the corresponding base quantities, since these are of great help in the teaching process; partly because they are in the curriculum, partly because they are in our everyday use.

We need to understand what each of these units are. It gives an excellent opportunity the teacher to make the best use of the Information Society. We can form groups, and give each group a task.

The groups are to look up what the units mean and how the definition changed in history. They can analyse, discuss and evaluate the information they gain from the internet. They also set up a short presentation for their mates to show what they found. These tasks are all active ways of learning according to Dale’s cone of experience [5].

We can help the groups’ work by giving hints and links they can begin their search from, like these:

- length, metre
https://en.wikipedia.org/wiki/History_of_the_metre
1793, prototype, 1983, Bay Zoltán
- mass, kilogram
<https://en.wikipedia.org/wiki/Kilogram>
1795, 1875, the “Grand k”, 1889, Watt Balance and Avogadro Projects
- time, second
<https://en.wikipedia.org/wiki/Second>
periodic phenomena, 1/86400, 1967, the seconds pendulum
- electric current and charge, amp and coulomb
<https://en.wikipedia.org/wiki/Ampere>
<https://en.wikipedia.org/wiki/Coulomb>
1881, 1946, $6.242 \cdot 10^{18}$, electron
- temperature, Kelvin and degree Celsius
<https://en.wikipedia.org/wiki/Kelvin>
<https://en.wikipedia.org/wiki/Celsius>
1743, 1956, ice-water-steam
- measure of substance and quantity, mol and piece
[https://en.wikipedia.org/wiki/Mole_\(unit\)](https://en.wikipedia.org/wiki/Mole_(unit))
1890, 1967, Avogadro, the ratio of one single entity
- luminous intensity, candela
<https://en.wikipedia.org/wiki/Candela>
1946, 1979, lumen, lux

We can note that the following link is very useful for each of the tasks: <http://www.si-units-explained.info/temperature/#.Wcd0sshJaUk>

Further tasks for the teacher:

- Once we have an agreement on the base unit we may find that these are either far too big or far too small for our project. Prefixes are used to solve this problem. We should also present these.
- It is important to strengthen the links between doctrines in public education. We should make remarks on orthography.
- Tamás Szabó Sipos made a series of cartoons to help the invention of SI at the time. It has an excellent sense of humour and the information the

cartoons use is correct. We can use for Hungarians only.

<https://www.youtube.com/watch?v=JyJBy24MIGw>

E) *The Record of Measurement*

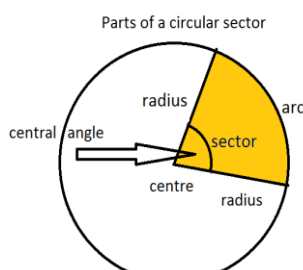
In engineering and in research it is important to make the documentation of our work. I worked out a type of documentation that helps our students best understand the steps of scientific cognition, and the analyzed problem itself also. In this part of the physics course I use two projects (and therefore work on two RoMs) to help my students to get familiar with the scientific method.

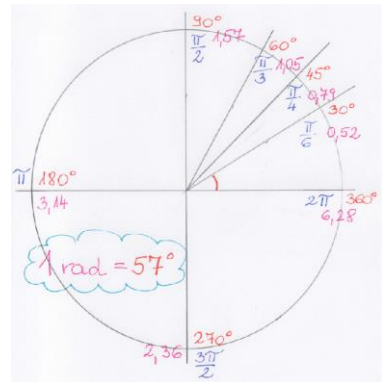
A project done together: Specifying angular measure

I find it significant to strengthen the connection between school subjects. This first project is dedicated to this idea: being also a teacher of mathematics, I have experienced that our students find it difficult to understand and work with it, although it emerges nearly every grade in the Hungarian National Curriculum.

As we progress in cognition we also understand what each part of the Record of Measurement stands for. Table II shows how the students get familiar with the method on an example. I do it as an outdoor activity if the weather allows.

TABLE II.
A ROM ON THE ANGULAR MEASURE PROJECT

RECORD OF MEASUREMENT	
Name (+mates):	
Venue:	
Date:	
Investigation: Measuring Angles Project	
Learn how angles can be measured by measuring length only	
The scientific background of the project:	
<ul style="list-style-type: none"> Roger Cotes in 1714 described this nature-based measure of angles. The term, radian first appeared in Belfast, at Queen's Collage in an exam paper, in 1873. Till 1995 it was an SI supplementary unit, now it is considered as a derived unit. In the figure basic notions of a circular sector are shown. 	
	
<ul style="list-style-type: none"> We define the measure of the central angle by the 	

<p>ratio of the length of the corresponding arc and the length of the radius in a circular sector.</p> $\alpha = \frac{\text{length of the arc}}{\text{length of the radius of the circle}} = \frac{a}{r}$ <ul style="list-style-type: none"> Another well-known measure of angles is degree. We measure it with a protractor. The full angle measures 360°. It is used in the study of circular motions. 	<p>Steps:</p> <ol style="list-style-type: none"> We draw circular sectors of different angles and radii. We will investigate 0°, 30°, 45°, 60°, 90°, 180°, 270° and 360°. The radii can vary from 10 to 100 cm. We measure the length if the arc using yarns, and the length of the radius using a tape measure. We divide the readings to get the angle in radian. $\alpha = \frac{\text{arc}}{\text{radius}}$ <ol style="list-style-type: none"> We repeat steps 2 and 3 for each sector 																																
<p>Tools: yarn, choke, tape measure</p>																																	
<p>Readings and analysis:</p> <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>angle</th> <th>30°</th> <th>45°</th> <th>60°</th> <th>90°</th> <th>180°</th> <th>270°</th> <th>360°</th> </tr> </thead> <tbody> <tr> <td>arc (cm)</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>radius (cm)</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>angle in SI</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		angle	30°	45°	60°	90°	180°	270°	360°	arc (cm)								radius (cm)								angle in SI							
angle	30°	45°	60°	90°	180°	270°	360°																										
arc (cm)																																	
radius (cm)																																	
angle in SI																																	
<p>Conclusion:</p> 																																	
<p>Notes:</p> <ul style="list-style-type: none"> We used the https://en.wikipedia.org/wiki/Radian site for our studies. The thickness of the choke's trace made it difficult to measure the length of the arc and radius precisely. Also, we made our readings from the tape measure to cm accuracy. 																																	

I have experienced in my practice that months later, when my students learn about this topic in math class, my colleagues are happy to hear about our project from the students. They regularly give the feedback that this project helped their work in the study of circles.

Getting started with statistics: The Cinderella project

Cinderella or The Little Glass Slipper is a fairy tale very popular all around the world. We know the story mostly in its version told in the Brothers Grimm in 1812. In the story, she needs to pick lentils out of ashes. In our project, the manual activity is very similar: we count pepper-grains.

The investigation is:

“How many pepper grains does a package contain?”

The project is a variation of the SI base quantity, the amount of substance. The students are encouraged to look back on what we have covered in the topic when focusing on the theory.

In this project, the students count how many grains there are in a package, using different methods. These methods may be

- a) counting one-by-one
- b) making groups of 10, counting the groups, and adding the rest
- c) making groups of 50, counting the groups, and adding the rest.

Very often they face the problem of not having the same result with the same method. This gives an excellent opportunity for the teacher to revisit what they have already discussed about errors. As the studied theory comes alive in a practical situation it consolidates the knowledge and demonstrates the duo of theory and its application. At this point knowledge goes to a thorough level of cognition, gets a deeper understanding.

As the groups are working they realize that their results are different. The difference of the groups’ results is worthy of a study. Counting how many grains there are in one particular package does not give the best answer to the question. Making simple statistics of the results of different packages is evidently a feasible solution. Thus, they also understand the need of cooperation in scientific research.

The class also needs to work out a table setting to logically show the results. In my practice, I found two appropriate solutions. These are in Table III and IV.

TABLE III.
NUMBER OF GRAINS IN PACKAGES

Groups	1.	2.	3.
method a)					
method b)					
method c)					
grains in the package					
average					

TABLE IV.
NUMBER OF GRAINS IN PACKAGES

Our package	method a)	method b)	method c)	our average

for more packages

Group							eventual average
average							

The students make a note on their average as the element of the set of averages, like: “*Seemingly in our package we had just about as many grains as all the other groups had I theirs.*”

When finishing the project, the issue of errors often occurs again. Namely some grains happen to “disappear” somehow: either some youngsters tend to eat them, or I find a few on the ground after collecting the equipment. The irreducible presence of error in a measurement turns into the focus again, thus vouching the previously so strange idea.

Using these data also gives an opportunity to practice statistical notions (like modus, median, mean, extent, etc.) for the math class.

III. CONCLUSION

Methodological research is in the service of better science education. Insertions of hands-on activities are present on the palette of classroom events. Also, our students need our help to evaluate and make the use if the Information Society for the best.

As an example, I planned and developed an optional chapter for the secondary school physics courses on Measurement Theory. I considered the use of the internet to motivate and engage in an active way our students.

ACKNOWLEDGMENTS

I express my gratitude to

- my PhD supervisor, Dr. Judit Illy (ELTE TTK),
- the Headmasters of my secondary schools: Mr. Károly Varsóci (Trefort Ágoston Bilingual Technical Secondary School, Budapest), and Mr. Csaba Mészáros (Madách Imre Gimnázium, Budapest),
- each of my students who participated in the development of the optional unit.

This study was funded by the Content Pedagogy Research Program of the Hungarian Academy of Sciences.

REFERENCES

[1] <https://www.ed.gov/stem>
 [2] Gupta Seema, Constructivism, as a paradigm for teaching and learning, In: International Journal of Physical and Social Sciences, 2011., Vol.:1, Issue 1., pp 23-47,
 [3] <https://www.youtube.com/watch?v=5qCT1aVmOIE>,
<https://www.youtube.com/watch?v=sqUVInhoz6s>,
<https://www.youtube.com/watch?v=t7Zr4A7yqUw>
 [4] G Gergely, M Menyhárd, A Sulyok, GT Orosz, B Lesiak, A Jablonski, J Tóth, D Varga: Surface excitation of selected polymers studied by EPES and REELS, Surface and Inteface Analysis, ISSN: 01422421, pp 1056-1059, 36/8, 2004
 [5] <http://teachernoella.weebly.com/dales-cone-of-experience.html>

Active contours in geoinformatics

K.Epresi, B. Gáti, Z. Tóth*

* University of Óbuda, Alba Regia Technical Faculty, Institute of Geoinformatics, Székesfehérvár, Hungary
toth.zoltan@amk.uni-obuda.hu

Abstract— Interesting tools of change management of geographic information systems may be the so-called active contour models, also called snakes, originally developed for image segmentation and movement recognition.

I. PROPERTIES OF ACTIVE CONTOURS

According to the original algorithm [Kass et al., 1988] an active contour is a curve defined by parameters, to which internal energy is attributed based on its shape, and external, potential energy originating from its environment. The energies are defined so as to minimize the energy of the contour along the edges we seek to find. The problems of seeking edges, segmentation and object detection are thereby transformed to an energy minimum problem. In the following, we will summarize the key properties of active contours based on the work of [Horváth, 2004].

The parametric equation of the active contour is:

$$v(s) = (x(s), y(s)) \quad (1)$$

The total energy attributed to the curve:

$$E_{teljes} = E_{belső} + E_{külső} \quad (2)$$

The internal energy is composed of two parts:

$$E_{belső} = E_{nyúlási} + E_{hajlító} \quad (3)$$

$$E_{nyúlási} = \frac{1}{2} \cdot \int_s \alpha(s) \cdot |v'(s)|^2 ds \quad (4)$$

$$E_{hajlító} = \frac{1}{2} \cdot \int_s \beta(s) \cdot |v''(s)|^2 ds \quad (5)$$

i.e. of the continuity of the contour and the smoothness of the contour, these regulating its shape. The $\alpha(s)$, $\beta(s)$ parameters in the formulae are weight functions (usually defined as constants).

$$E_{külső} = \int_s E_{Kép}(v(s)) ds \quad (6)$$

The external energy is calculated from the image, and it may be defined in a number of ways. The value of $E_{Kép}$ as the simplest possible solution (e-g in a binary image):

$$E_{Kép} = I(x, y), \text{ and} \quad (7)$$

$$E_{Kép} = G_\sigma(x, y) \cdot I(x, y) \quad (8)$$

Where $I(x, y)$ are the intensity values of the image, and G_σ is the two-dimensional Gaussian kernel with standard deviation σ . For a more general case, e.g. an 8-bit grayscale colour depth image, the following definition may be used for the external energy:

$$E_{Kép} = -|\nabla I(x, y)|^2, \text{ and} \quad (9)$$

$$E_{Kép} = -|\nabla G_\sigma(x, y) \cdot I(x, y)|^2 \quad (10)$$

Where $I(x, y)$ are the intensity values of the image, G_σ is the two-dimensional Gaussian probability density function and ∇ is the symbol of the gradient operator.

For all solutions, the value of $E_{Kép}$ is minimal at the valuable areas (such as the boundaries of objects); this condition must evidently be satisfied by some type of image pre-processing (highlighting outlines). The external energy thus defined shifts the contour to the valuable locations (of low potential energy), while minimising the total energy of the contour expressed based on the information so far as:

$$E_{teljes} = \int_s \frac{1}{2} \cdot (\alpha(s)|v'(s)|^2 + \beta(s)|v''(s)|^2) + E_{Kép}(v(s)) ds \rightarrow \min \quad (11)$$

While the contour moves towards the lower energy parts of the image in order to satisfy the above equation, it may not freely take up any shape, hence the role of the members describing the internal energy: it may not elongate (the polygon points may not be distanced from one another), and the curve may not bend to an unlimited

extent (the change of curvature is limited in comparison to the initial state).

The contours that minimise the previous equation, must satisfy the following Euler-equation:

$$\alpha \cdot x''(s) - \beta x''''(s) - \nabla E_{Kép} = 0 \quad (12)$$

In practice, active contours are naturally not continuous curves, but discrete polygonal lines:

$$v_i = (x_i, y_i) \quad (13)$$

where $i = 1 \dots N$, N is the number of polygonal points describing the contour.

The description of the internal energy components comply with the discretisation used in image processing. The difference quotient in the elongation energy component may be calculated in practice as follows [Williams et al. 1992]:

$$\left| \frac{dv(s)}{ds} \right|^2 \approx |v_i - v_{i-1}|^2 = (x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 \quad (14)$$

while in the energy component of bending:

$$\left| \frac{d^2v(s)}{ds^2} \right|^2 \approx |v_{i-1} - 2 \cdot v_i + v_{i+1}|^2 = (x_{i-1} - 2 \cdot x_i + x_{i+1})^2 + (y_{i-1} - 2 \cdot y_i + y_{i+1})^2 \quad (15)$$

The value of the $\alpha(s), \beta(s)$ weight parameters in (11) may be empirically assumed as constants in a particular example for the entire contour, or a different value may be attributed to each polygonal point.

In the course of the iteration, a contour containing the initial number of vertices may be used, and new vertices may also be introduced if the task requires if two consecutive vertices were to be excessively distanced from one another, and if the curvature defined by three points were to change significantly with the iteration steps. Theoretical possibilities of the opposite case also exist, i.e. if a point may appear to be a line point during the iteration, it may be omitted from the subsequent calculations. Conditions may be set for inserting and omitting points, depending on the task in question; one such natural condition may be that the distance between points may not be smaller than the geometric resolution of the image.

II. QUESTIONS OF USING ACTIVE CONTOURS IN GEOINFORMATICS

When examining application possibilities of active contours in geoinformatics, their use in secondary data collection (digitisation) goes without saying. In the following, the initialisation and parameterisation possibilities of contours will be shown using the example of a cadastral map detail. The test has been run using an application prepared for Matlab mathematical developers' environment [Tomazevic et al., 2002]. The cadastral map detail shows and U shaped building. (Figure 1.) In the figure, we have indicated the smoothed images prepared using equation (8) for values of $\sigma = 2, \sigma = 5$, and in Figure 2., we have shown the external energy field corresponding to the images. (σ two-dimensional Gaussian with standard deviation σ .) It can be seen from the figures that by increasing σ our area of access increases, i.e. the image detail over which the initiated contour may converge into the original contours.

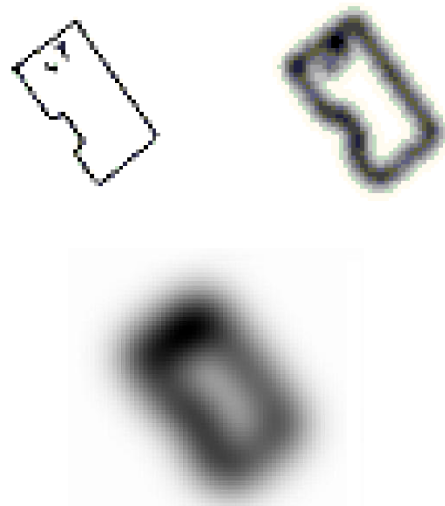


Figure 1. Smoothed versions of the original image as a function of deviation ($\sigma = 0, \sigma = 2, \sigma = 5$)

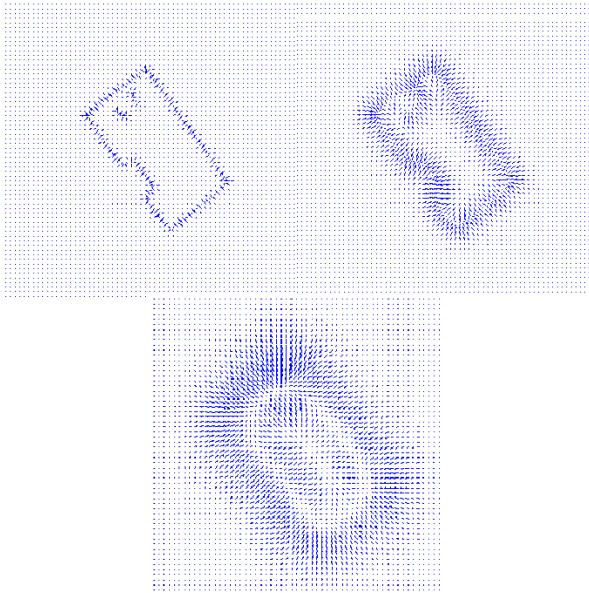


Figure 2. The “energy field” corresponding to the images ($\sigma = 0, \sigma = 2, \sigma = 5$)

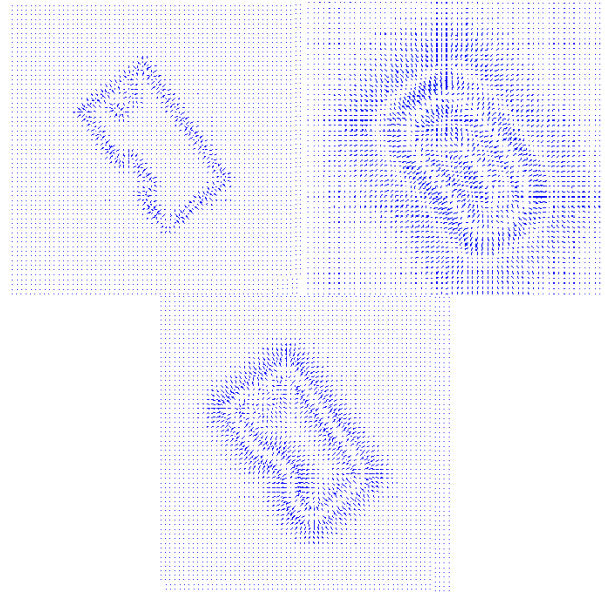


Figure 4. The “energy field” corresponding to the smoothed gradient images ($\sigma = 0, \sigma = 2, \sigma = 5$)

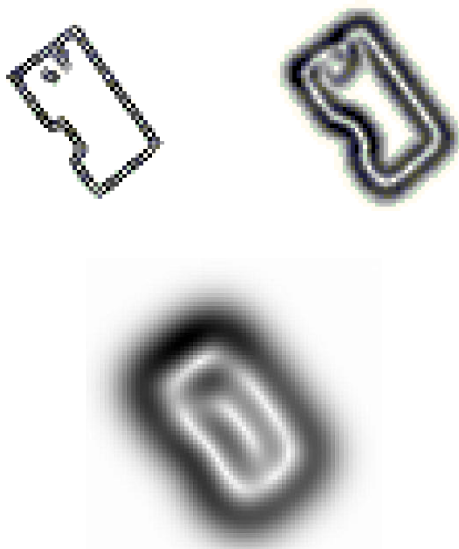


Figure 3. Gradient images ($\sigma = 0, \sigma = 2, \sigma = 5$)

The disadvantage of smoothing is however observable: the value of the σ parameter may not be increased freely, because the increase of the area of access is also accompanied by the disappearance of the edges to be detected. In our case, for example, the U shaped indentation disappears. (Figure 1.) Similarly, by calculating the gradient vectors and using equation (10), the external energy field may be produced, as shown in figures 3 and 4. To increase the access area, the σ parameter may be increased with similar constraints as if we were to start from the intensity values. For conventional contours, all authors [Horváth, 2004] mentions, besides the calculation problems of the access area mentioned above, the difficulties associated with the detection of concave shapes. A good example of this – from the field of cadastral maps – may be the case of U shaped buildings as shown above.

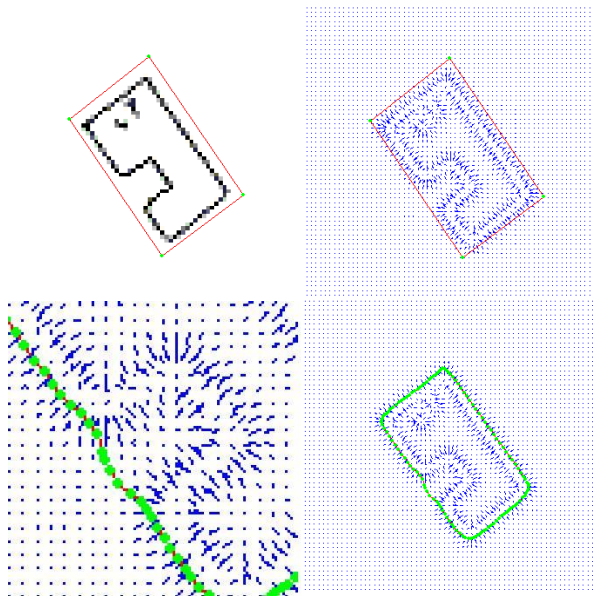


Figure 5. Example for the segmentation problem of concave shapes: (a) original image, (b) calculated gradient image, (c) final result of segmentation, (d) the concave part of the building magnified

The cut-out of the building with considerable concavity from Figure 5. (a), with the initialisation contour, the following figure (b) showing the external energy field calculated from the gradient image of the original image with a value of $\sigma=1$, on which the initial position is also indicated. On figure (c) the equilibrium position of the contour is indicated in green. It can be observed that while the majority of the building has been segmented appropriately, this was not so for its concave part. The causes are revealed by figure (d): there is no gradient vector that would “pull in” the curve into this area, as there are quasi-identical gradient vectors in this part of the image pointing in the opposite direction, i.e. towards the sides of the building, while there is nothing in the middle to draw the contours inside.

The problems mentioned above, i.e. the difficulty of initialisation and the problem of too large shapes naturally cause problems in other application areas of image processing; one possible solution is using the Gradient Vector Flow [Xu et al., 1998] vector field. The vector fields of the Gradient Vector Flow are defined as follows [Horváth, 2004]:

$$v(x, y) = [u(x, y), v(x, y)] \quad (16)$$

This vector field must minimise the following energy function [Horváth, 2004]:

$$\varepsilon = \iint \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |v - \nabla f|^2 dx dy \quad (17)$$

where f is the contour image derived from the original image, and μ is a weight parameter. The GVF field is obtained by solving the following Eulerian system of equations [Horváth, 2004]:

$$\mu \nabla^2 u - (u - f_x)(f_x^2 + f_y^2) = 0 \quad (18)$$

$$\mu \nabla^2 v - (v - f_y)(f_x^2 + f_y^2) = 0 \quad (19)$$

Where ∇^2 is the Laplace operator. Figure 6 shows an example of the external energy field thus calculated for Figure 1. (a). Figure 6. (a) shows the field and the place of initialisation, with (b) showing the running results after ~20 iterations. The cause of the successful segmentation is shown in Figure 6. (c): the vectors of the GVF field point to the concave part of the building, as expected.

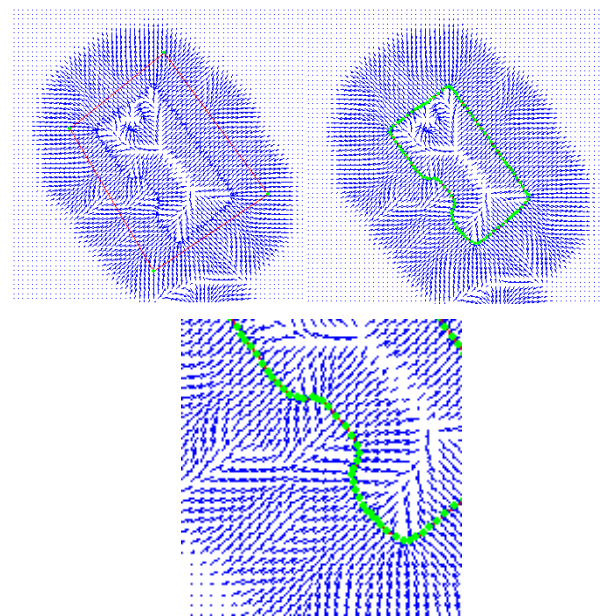


Figure 6. GVF field

In our article, we have introduced the properties of active contours, with application possibilities in geometric change management shown through practical examples. In this latter case, the necessity of initialisation does not emerge as a disadvantage, as the previous state may be taken as the first iteration step.

REFERENCES

- [1] Kass, M. – Witkin, A. – Terzopoulos, D.(1988) Snakes: Active Contour Models, International Journal of Computer Vision, Vol. 1, pp. 321-331.
- [2] Horváth, P. (2004): Képek szegmentálása szín és mozgás alapján. Diplomamunka, Szegedi Tudományegyetem, p.61.
- [3] Williams, D. – Shah M.(1992): A Fast Algorithm for Active Contours and Curvature Estimation. CVGIP:Image understanding Vol 55. pp. 14-26.
- [4] Tomazevic, D. (2002): GVF Snake demo. (Matlab alkalmazás)
- [5] Xu C. – Prince, L. J. (1998):Snakes, Shapes, and Gradient vector Flow, IEEE Transactions on Image Processing, Vol. 7, pp. 359-

Network security in the age of SDN, AI and BIG DATA

Péter Bálint

Széchenyi István Egyetem

Győr, Hungary

Email: peter.balint4@gmail.com

Abstract— The increasing number of security attacks drives the IT security society to find answers to the new challenges. In the last few years many new technologies appeared, the lifecycle of which achieved the level of usability in complex IT security cases. The most relevant technologies and trends are Big Data, Cloud and SDN. Big Data covers many different disciplines, data warehouse design, data engineering, data analysis, data driven decision making, data driven artificial intelligence and deep learning. From these, the most relevant are the artificial intelligence based Intrusion detection systems. Cloud computing, often referred to as “the Cloud,” has been another hot paradigm in IT over the last few years. Its popularity based on its scalability and flexibility, which are key factors in the fast changing business environment. The most important idea behind it is that the cloud enables companies to consume a compute resource, such as a virtual machine, storage or an application, as a utility, rather than having to build and maintain on premise infrastructure. SDN (Software-Defined Networking) is a new trend in computer networking, where the control plane and the data plane are separated. SDN can make the networks much more efficient and scalable. These new technologies are introduced in this paper from the point of view of possible usage in security cases. Then a reference architecture is presented, which is built up from these elements. Next an IDS solution are introduced and the possible future challenges and tasks are summarized. Finally, our results and future development possibilities are summarized.

Keywords— *IT security, IDS, SDN, Big Data, Artificial Intelligence, Machine learning*

I. INTRODUCTION

In the last years, computers became an important part of our life. The bigger usage time and quantity of different software solutions increased possibility of attacks. The increasing trend showed as that the currently used security software and technics are not enough. On other side, the computing capacity and the increasing intelligence of our IT infrastructure gave us new tools to protect our IT environment and data from attacks. In this paper we would like to cover these topics. At the beginning we start with a brief summary of new technologies, this is followed by current IDS solution introduction, this topic separated in two chapters. First is a general summary, the second focus to the increasing usage of Artificial intelligence in IDS implementations. Then we

introduce Cloud computing as a very popular technology. Last but not least we propose a solution that builds up from the introduced technologies. This solution could provide enhanced IDS functionality in service provider and big enterprise segment.

II. SDN

Nowadays the networks flexibility and cost are contradictory with each other. Managing and operating the networks has become very difficult and complex activity. The technical answer to these problems is Software Defined Networking (SDN). The Software Defined Networking brings a very different mindset in networking as before. It provides flexible control over network policies and traffic flows. It achieves this through different network equipment. We speak now about SDN network and not about standalone network appliances. By using Software Defined Networking the changes in network could be implemented centrally and

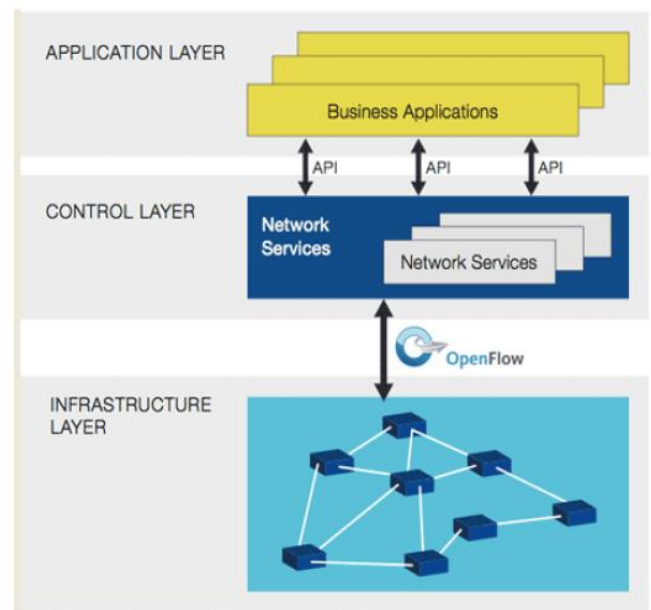


Fig. 1. SDN Architecture [1]

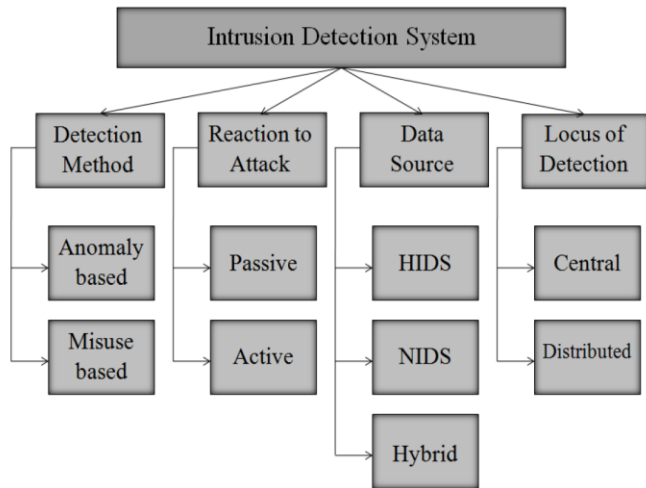


Fig. 2. IDS categorization [2]

automatically.

The most important features of SDN:

1. Centralization of network operations and management
2. Automatic traffic management possibilities
3. Easy customization

The SDN Architecture has three layers, see Fig. 1:

- Infrastructure layer: The lowest layer, where traffic forwarding is happening.
- A middle layer, SDN controller layer
- Application layer which consists of the applications and services.

III. IDS

With the increasing number of IT services, the intrusion detection systems (IDS) have become important tools for ensuring security. IDS is a device collect information from a variety of network sources using intrusion detection sensors, and analyze the information for signs of intrusions that attempt to compromise the confidentiality and integrity of networks. There are many types of IDS, see Fig. 2

An IDS builds up from several components:

- Sensors that generate security events.
- Console to monitor events, alerts and control the sensors.
- Central Engine, what records events logged by the sensors in a database and uses a system of rules to generate alerts from security events received.

The Functions of IDS:

- Monitoring and analyzing the computers, users, network traffics and systems
- Analyze system configuration pattern and vulnerabilities.
- Assessing systems and file integrity
- Tracking the user policy abuses.

Classification based on data source:

- Network based IDS (NIDS): These are installed at special points inside the network in order to monitor all traffic on the network.
- Host based IDS (HIDS): These operate on individual devices or hosts on the system. This will monitor all the incoming and outgoing packets on the device only and can notify the administrator or user of any suspicious activity.
- *Hybrid IDS*: Is the mixture of NIDS and HIDS. Therefore, it can get the broadest scope.

Detection method:

- Misuse Detection: Intrusions are detected by matching actual behavior recorded in audit trails with known suspicious patterns. While misuse detection is fully effective in uncovering known attacks, it is useless when faced with unknown or novel forms of attacks for which the signatures are not yet available. Moreover, for known attacks, defining a signature that encompasses all possible variations of the attack is difficult. Any mistakes in the definition of these signatures will increase the false alarm rate and decrease the effectiveness of the detection technique. It begins protecting the computer/network immediately upon installation. But the major drawback of misuse-based detection is that it requires frequently signature updates to keep the signature database up-to-date. Misuse detection system use various techniques including rule-based expert systems, model-based reasoning systems, state transition analysis, genetic algorithms, fuzzy logic, and keystroke monitoring
- Anomaly Detection: Different from misuse detection, anomaly detection is dedicated to establishing normal activity profiles for the system. It based on the assumption that all intrusive activities are necessarily anomalous. Anomaly detection studies start by forming an opinion on what the normal attributes for the observed objects are, and then decide what kinds of activities should be flagged as intrusions and how to make such particular decisions. Anomaly detection approaches often require extensive learning process in order to characterize normal behavior patterns. Unfortunately, the early IDS solutions what based on anomaly detection often produce a large number of false alarms, because the normal patterns of user and system behavior can vary wildly. Modern enterprise network environments amplify this disadvantage due to the massive amounts of dynamic and diverse data analysis. Despite this shortcoming, researchers assert that IDS based on anomaly detection are able to detect new attack forms. Various data mining algorithms could be used for anomaly detection, including statistical analysis, sequence analysis, neural networks, artificial intelligence, machine learning, and artificial immune system.

Actual	Predicted	
	Normal	Attack
Normal	True negative (TN)	False positive (FP)
Attack	False negative (FN)	True positive (TP)

Fig. 3. IDS outcomes[2]

Before we go to next section, we summarize the main categories of attack types:

- DoS attack: Denial of Service attack results by preventing legitimate requests to a network resource by consuming the bandwidth or by overloading computational resources.
- Probing attack: These attacks collect information about target system prior to initiating an attack.
- User to Root (U2R) attack: In this case, an attacker starts out with access to a normal user account on the system and is able to exploit the system vulnerabilities to gain root access to the system.
- Root to Local (R2L) attack: In this case, an attacker who doesn't have an account on a remote machine sends packet to that machine over a network and exploits some vulnerabilities to gain local access as a user of that machine

In the next section we briefly sum up the different measurements of IDS systems. Fig. 3 depicts the four possible outcomes of IDS from a given event and prediction. The four outcomes are representing as True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP). True negatives (TN) are events, which are actually normal and successfully labeled as normal when there is no intrusion attack. True positive (TP) are events which are actually attack and successfully labeled as attack when there is any intrusion attack. These two events correspond to correct operation of the IDS. False positive (FP) are events which are attack in the normal event and correctly showed as attack. False negative (FN) are events which are attack in the attack event but it showed wrongly as normal.

IV. DATA MINING BASED IDS

Variety of Data mining technics used in IDS systems. An IDS monitors network traffic in a computer network like a network sniffer and collects network logs. Then the collected network logs are analyzed for rule violations by data mining algorithms. When any rule violations are detected, the IDS alert the network security administrator or the automated intrusion prevention system (IPS). The generic architectural model of data mining based IDS is shown in Fig 4.

- Audit data collection: IDS collect audit data and analyze them by the data mining algorithms to detect suspicious activities or intrusions. The source of the data can be host/network activity logs, command-based logs, and application-based logs.
- Audit data storage: IDS store the audit data for future reference. The volume of audit data is extremely large. Currently adaptive intrusion detection aims to solve the problems of analyzing the huge volumes of audit data and realizing performance optimization of detection rules.
- Processing component: The processing block is the heart of IDS. It contains the data mining algorithms that responsible of detection of suspicious activities. Algorithms for the analysis and detection of intrusions have been traditionally classified into two categories: misuse (or signature) detection, and anomaly detection.
- Reference data: The reference data stores information about known attacks or profiles of normal behaviors.
- Processing data: The processing element must frequently store intermediate results such as information about partially fulfilled intrusion signatures.
- Alert: The output of IDS that notifies the network security officer or the automated intrusion prevention system (IPS).
- System security officer or intrusion prevention system (IPS) carries out the prescriptions controlled by the IDS.

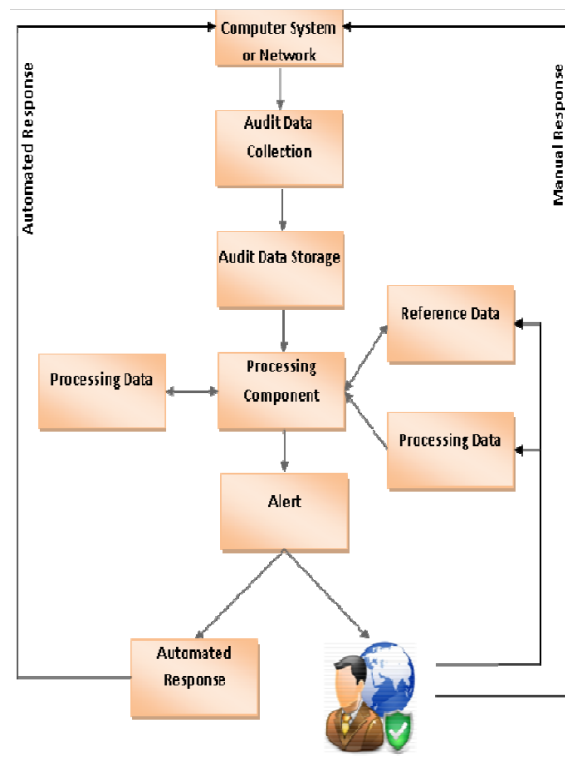


Fig. 4. Flowchart of data mining based IDS [3]

There are many type of algorithm available for analyzing the datasets like decision tree, naïve Bayesian classifier, neural network, Support Vector Machines, and fuzzy classification, etc. Data mining based intrusion detection algorithms aim to solve the problems of analyzing the huge volumes of audit data and realizing performance optimization of detection rules. However, there are still some drawbacks in currently available commercial IDS, such as low detection accuracy, large number of false positives, unbalanced detection rates for different types of intrusions, long response time, and redundant input attributes. The database of logs is complex, dynamic and contains many different attributes. The problem is that not all of these attributes may be needed to build efficient and effective IDS. The use of redundant attributes may interfere with the correct completion of mining task, because the information they added is contained in other attributes. The use of all attributes may simply increase the overall complexity of detection model, increase computational time, and decrease the detection accuracy of the intrusion detection algorithms. Therefore, beside the algorithm, the relevant data selection is a key element of reduction of poor detection accuracy [3]. Moreover, the big quantity of data strongly reduces the IDS performance. Data mining methods could be used for selecting relevant data, some possible method: principal component analysis (PCA), genetic search, and classifier ensemble methods. A good solution was the [4] wrapper-based feature selection algorithm. This found the relevant features from the dataset and the random mutation hill

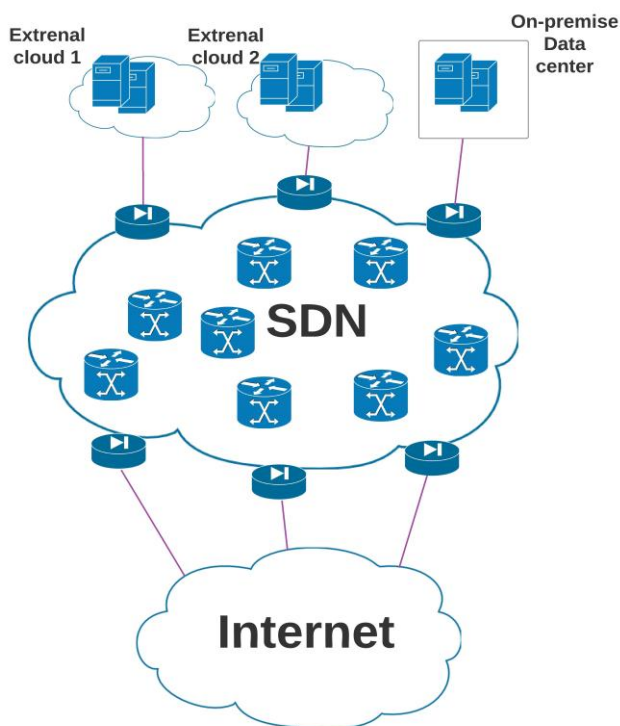


Fig. 6. Proposed system architecture

TABLE I. MAX VARIATION OF TIME BETWEEN PACKETS (MILLISECONDS) [3]

	SVM	NN	GA	Proposed Algorithm
Normal	99.4	99.6	99.3	99.93
Probe	89.2	92.7	98.46	99.84
DoS	94.7	97.5	99.57	99.91
U2R	71.4	48	99.22	99.47
R2L	87.2	98	98.54	99.63

climbing method, and the linear support vector machine (SVM) provide good results. Dewan [3] used ID3 algorithm based decision tree for selecting the relevant attributes from dataset. The ID3 algorithm constructs decision tree using information theory, which choose splitting attributes from the training dataset with maximum information gain. Information gain is the amount of information associated with an attribute value that is related to the probability of occurrence. Entropy is the quantify information that is used to measure the amount of randomness from a dataset. When all data in a set belong to a single class, there is certainty then the entropy is zero. The objective of ID3 algorithm is to iteratively partitioning the given dataset into sub-datasets, where all the instances in each final subset belong to the same class. Naïve Bayesian tree (NBTree) is a hybrid learning approach of decision tree and naïve Bayesian classifier. In NBTree, nodes contain and split as regular decision-trees, but the leaves are replaced by naïve Bayesian classifier, the advantage of both decision tree and naïve Bayes can be utilized simultaneously. Depending on the precise nature of the probability model, NB classifier can be trained very efficiently in a supervised learning. The decision tree-based attribute weighting with naïve Bayesian tree are suitable for analyzing large number of network logs. The comparison to other algorithms we could see Table 1. As we could see the proposed algorithm give better performance, this highlights a good combination of algorithm and setting perform a good level of intrusion detection. [3]

V. CLOUD

The Cloud (Cloud Computing) as the new popular trend in IT. The popularity based on the scalability and flexibility that are key factors in the fast changing business environment. The main idea that the cloud enables for companies to consume a compute resource, such as a virtual machine, storage or an application, as a utility. Rather than having to build and maintain on-premise infrastructure. The cloud is relevant in two cases to our topic.

Cloud as resource

Cloud is a very scalable and pay-as-you use service, so in cases when company need extra computing resource for analyzing logs. It could be quite useful.

Cloud as a new part of infrastructure where security needed.

The increasing number of cloud based IT infrastructures and hybrid cloud network bring new challenges in security. The IDS is an essential part of on-premise enterprise

networks, but if we migrate to cloud IDS solutions are also required to ensure the security.

VI. PROPOSED SOLUTION

After the brief summary, now we would like to introduce a possible architecture that builds up from above mentioned technologies. The architecture could be usable at big enterprises and service provider segment. The architecture builds up from an SDN based IT network, on-premise data center and two external datacenter (two cloud provider, multi-cloud), see Fig. 6. At the edge of these elements, there are network traffic analyzers, which collect and forward all relevant information to IDS analyzer. The IDS analyzer engine run in the on-premise datacenter, but in special cases the computing capacity could be increased with an external cloud provided computing resource. The architecture without the IDS related elements could be a common service provider or big enterprise architecture. In these segments because of the big traffic load is very difficult and expensive to implement an IDS system that cover the full network. In the next section I propose an which make available the usage of IDS system in service provider or big enterprise segment.

The introduced architecture has different subsystems that have their on controllers. The subsystems:

- SDN: main task is the traffic forwarding among the network elements. SDN controller is the brain it is responsible for controlling of network.
- IDS: collect and analyze the network traffic
- Cloud and on-premise datacenter

The available IDS solution could provide only limited capacity that is not on the field of service provider segments requirements. To make available the IDS functionality at this size of traffic we suggest a multi-layer intrusion detection system. In this model the first layer not analyze all packet in

detail it investigate only the traffic patterns and if it detect something abnormal then investigate the traffic deeper, see Fig. 7.:

1. The network packet logger collects the packets and forwards the relevant information to the data analysis system.
2. The analyzer could identify abnormal traffic based on the historical data.
3. In abnormal case the SDN controller get informed about it and it change the path of traffic to go through the intrusion detector
4. The Intrusion detector analyzes the traffic deeper and intervenes if it is reasonable.

Most parts of this architecture are available and accessible in open source format. However, some of them should be designed a developed. The IDS system use Snort a lightweight open source IDS for packet logging and for Intrusion detection. Snort is a Signature based IDS but it is possible to add new rules and increase the detection accuracy with artificial intelligence logic. As we showed IDS introduction chapter the AI supported IDS solution could provide very good results.

The analyzer is a system in the system it builds up a distributed storage for storing the big quantity of data and data analysis and artificial intelligence logic for learning the normal workload and detect anomaly.

This solution make available the usage of IDS in high load of traffic cases, so in service provider or big enterprise networks. However, as usual there are limitations too. In the first layer the system investigate only the traffic patterns, so the system could only detect the attack which detectable with this method. However, in other cases the attacks easily go through on the network. Therefore, in cases when solutions or customers require more detailed intrusion detection and prevention then the traffic should go through the AI supported

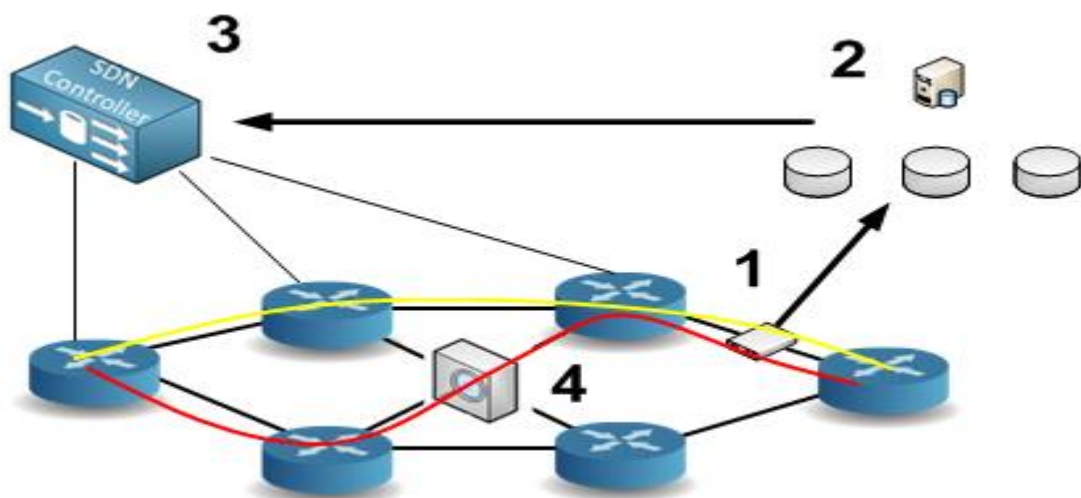


Fig. 7. IDS outcomes[2]

IDS in every case. We could increase the first layer IDS detection performance if we integrate the cloud and on-premise datacenter controller to IDS system. With this solution, the IDS could monitor the services parameters like CPU or memory usage. These values could help to detect attacks. This solution makes available the usage of IDS solutions in service provider segment and so increase the network resistance again attacks.

VII. FUTURE

As we mentioned many parts of the solution is available, but some of them should be designed and developed. First important task is the design and development of the IDS systems analyzer part. The architecture of the analyzer should builds up from:

- Distributed file system.
- Management unit that analyze the historical and input information and it is able to recognize anomaly.

The Fig. 8. illustrates the architecture and the connections with SDN and Cloud controller elements.

VIII. CONCLUSION

In big enterprise and service provider segment the IDS functionality that cover the whole network and all traffic is expensive because of the huge costs. One goal of this paper was the introduction of cutting edge networking technologies, which could make available the IDS functionality in these segments. The other important goal was to propose a multi-layer IDS solution, which could make available the IDS functionality in service provider segment. The introduced solution increases the security level of service provider networks. Main parts of this system are available, but some of them should be designed and developed. This need generate the topic of the future work, the design and development of analyzer element, which is part of IDS system.

So as summary, we achieved in the Abstract specified goal we brought forward the most relevant technologies and

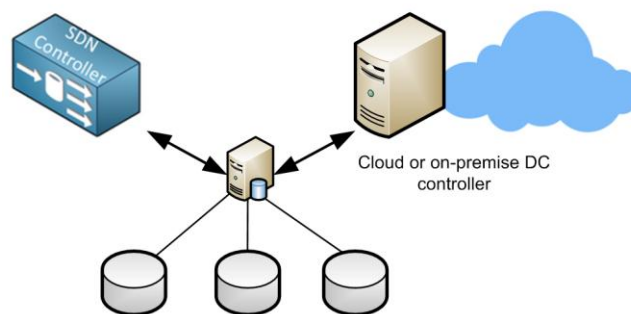


Fig. 8. High level architecture of IDS systems analyzer and interfaces with other network controller

proposed an IDS system and architecture that makes available the IDS functionality.

REFERENCES

- [1] S Bibi , B Tariq , A Batool ,F Mustafa, "SECURITY ANALYSIS OF SOFTWARE DEFINED NETWORKING (SDN)", International journal of Computer Science & Network Solutions, Jul.2015-Volume 3.No.7, ISSN 2345-3397
- [2] A. Lourdu Caroline , K. Ramanan, A. Arul Sophia, V. Vennila, "Revealing IDS through Data Mining", *International Journal of Computing Algorithm*, vol. 03, May 2014, Pages: 936-939
- [3] D. Md. Farid, J. Darmont, M Zahidur Rahman "Attribute Weighting with Adaptive NBTree for Reducing False Positives in Intrusion Detection", *International Journal of Computer Science and Information Security*, Vol. 8, No. 1, 2010, ISSN 1947-5500
- [4] Y. Li, J.L. Wang, Z.H. Tian, T.B. Lu, and C. Young, "Building lightweight intrusion detection system using wrapper-based feature selection mechanisms," *Computer & Security*, Vol. 28, Issue 6, September 2009, pp. 466-475.
- [5] A Zarrabi, A Zarrabi "Internet Intrusion Detection System Service in a Cloud" *International Journal of Computer Science*, Vol. 9, Issue 5, No 2, September 2012, ISSN: 1694-0814

Open source software application in point cloud processing

B. Gáti

*University of Óbuda, Alba Regia Technical Faculty, Institute of Geoinformatics, Székesfehérvár, Hungary
gatibence@gmail.com

Summary: In the dissertation we're giving an example on the potential of processing data with open source code softwares in relation to the survey of a Roman Catholic monumental church with a laser scanner. During the processing we created models of the building layout and the sections of the church which will hopefully reveal a connection between the length of church walls and interior heights, and fathom units used in feudal Hungary. Furthermore we're pointing the pros and cons of using different data sources.

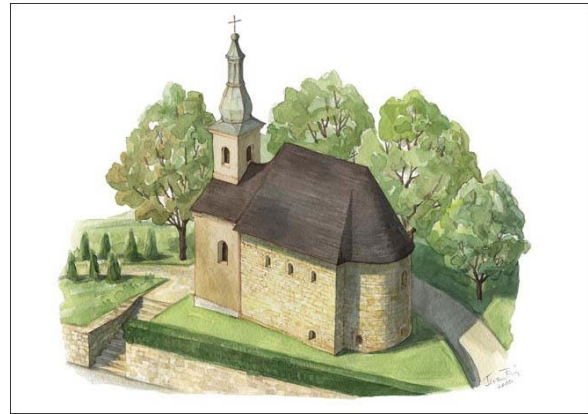


Figure 1. Graphic of the church

1. THE CHURCH OF TARNASZENTMÁRIA

One of Hungary's oldest yet still functional churches are in Tarnaszentmária. What is also interesting about it is that according different statements from Hungarian and Czechoslovak historians it is likely that it was built during the Slavic era of our country, from before the Hungarian Conquest with the intent of holding royal burial ceremonies. According to Tibor Gerevich the church must have been built in the middle of the XII. century. However, József Csemegi suggests the nave must have been built sometime during the IX-X. century and is backed by the fact that the building has features originating from Early Christian and Byzantine era.

Such features are for example the ornate berms at the feet of thin columns. In his opinion "the nave of the church in Tarnaszentmária is the gathering place for the pagan Aba nation, whose ornaments were brought along with our ancestors after their secession from the Khazar Empire." After further archeological discoveries it became clear that the sanctum and the nave were built at the same time. "In the nave of the small church a tomb was discovered with similar features to the building and in its sanctum a "pit-grave" was located keeping the previously disturbed bones of someone who was once important. The latter discovery makes it obvious that the church that means so much to us originally wasn't a gathering place for pagans but it was built as the burial site for a once important family -at least partially, along with serving other functions- before the birth of Roman era architecture in Hungary." As for who built it, either of two people seem likely: Grand Prince Géza and his brother, Mihály.

So the church, just like many other similarly built churches, is suitable to serve as basis to our inspection of the royal fathom.



Figure 2. The church of Tárnaszentmária

2. SURVEY TECHNOLOGIES

We planned on using three different technologies for the survey. UAV, terrestrial photogrammetry, laser scanner, and the traditional measurement using a total station to serve as a basis. By using unmanned aerial vehicle we planned on creating a 3D model and a pointcloud, however the takeoff was foiled by stormy weather, so we were unable to take aerial pictures.

The pictures used for terrestrial photogrammetry were taken by two kinds of digital cameras. The Nikon L340 uses autofocus and takes pictures with 20-megapixel resolution while pictures taken with the Sony Alpha have a 12 megapixel resolution and has a fixed focal length.



Figure 3. Sony Alpha DSLR A350 and Nikon L340

We also used two different laser scanners for terrestrial scanning, a FARO and a LEICA C10 type scanner, the article includes only the data from the LEICA C10 scanner. Measurements with using a total station were made to define the coordinates of control points in a local system.



Figure 4. Surveying using a laserscanner

3. PROCESSING

The processing of laser scanner-based measurements began with the usage of a commercial software called Cyclone by Leica. The evaluations of the data from other surveys were based on the pointcloud resulting from the Cyclone based processing. The software connects pointclouds measured from different stations with 3D congruency transformation. The residual error of the transformation originate from the GCP and -to be more cost-efficient- simple black and white marks printed on paper and are around 1-2mm.

The need for the use of freeware softwares used in education and research in addition to commercial software came up during the processing of laser scanning. We decided with the use of a software called CloudCompare which is used to process and display pointclouds and whose many functions make it suitable for tasks such as this and can be further developed as needed. With the software we are able to join pointclouds with different methods (eg. ICP algorithm, Helmert transformation) Since in our case there's little overlay among measurements made from different stations -to calculate transformation parameters from the co-ordinates of control points- we created an application for its latter function using octave. The program calculates the parameters of rigid body transformation in a way that is compatible with the pointcloud processing software.

With the application coordinates of certain stations can be transformed into a local system defined by measurements using a total station.

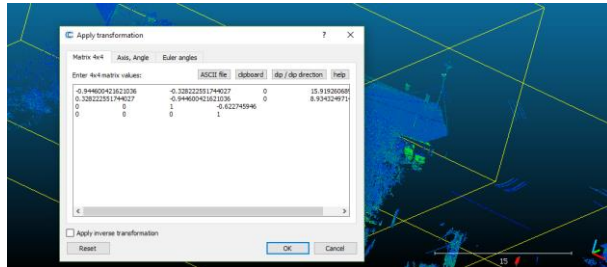


Figure 5. Transformation of each station

The next step was the cleanup of the pointcloud, we manually removed certain points, for example vegetation appearing as noise. The resulting pointcloud only contained points of the church (Figure 6). Using the intersection function of the program sections or the layout of the church can be easily manufactured and exported to any CAD software to serve as basis for any further research concerning royal fathom.



Figure 6. The upper and lower reaches of the church in a single pointcloud

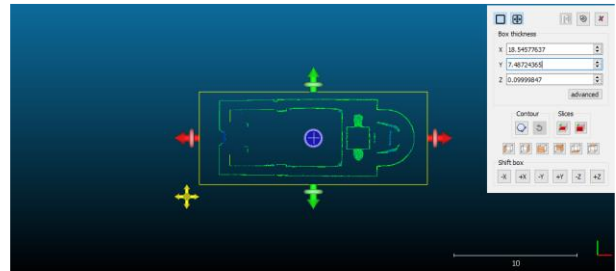


Figure 7. Editing the layout

As mentioned in the introduction the pointclouds weren't only surveyed via laser-scanner but also by terrestrial photogrammetry using non-metric cameras. Our goal was to compare the results of using different technologies from different points of view (difference in cost, time, reliability). During the processing we used a general-purpose commercial software (Photoscan) and later we thought it would be expedient to use other, open source programs. During the evaluation, we chose well identifiable points to serve as control points.

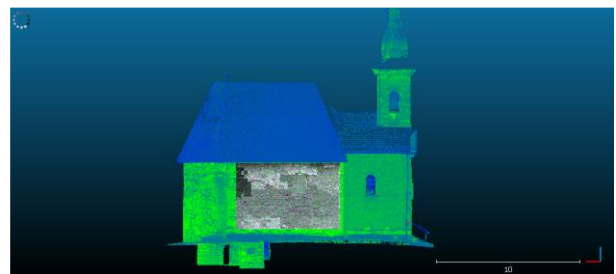


Figure 8. Pointclouds from different sources layered on each other

We designated a sample area to examine the accuracy of the pointcloud created with photogrammetric evaluation (Figure 8.). To compare the two methods we used Cloudcompre's "pointcloud distance" function to calculate the average distance of the laser-scanned pointcloud and the pointcloud resulting photogrammetric evaluation on the sample area with the former serving as reference. The result was ~0.6cm

4. OTHER RESULTS

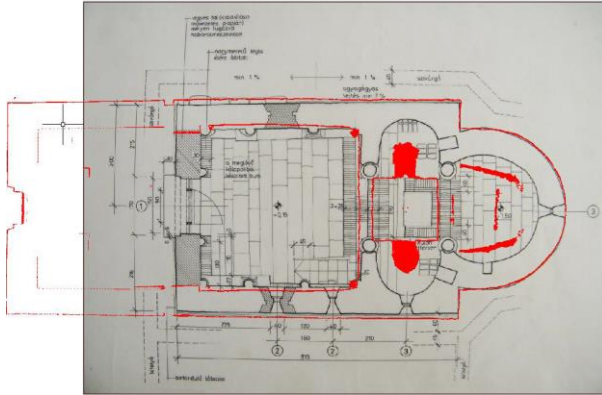


Figure 9. Floor plan in new and old technology

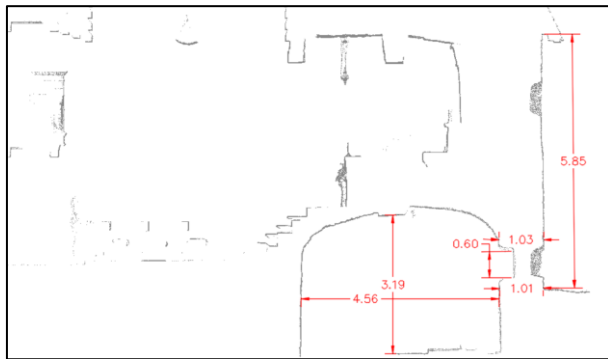


Figure 10. Vertical intersect with sizes

5. SUMMARY

Based on our findings it can be said that both technologies are more than suitable for the architectural survey mentioned in the article. The resulting pointclouds are sufficient to serve as basis for further examinations (creating new layouts, sections or simply measuring length) without any further field survey. By comparing the two the cost-efficiency of photogrammetry is worth mentioning along with almost equally reliable results. As a drawback, it may require additional survey (eg. defining control points).

The research was sponsored by the “Emberi Erőforrások Minisztériuma - Új Nemzeti Kiválóság Program”.

REFERENCES

- [1] Busics György: A forgotten length measurement, the royal fathom and its connection to the sizes of medieval buildings. Study, Székesfehérvár, 2015. 110 pages
- [2] John Ashburner, Karl J. Friston: Rigid Body Registration. London, 2004.
- [3] Kozák Károly: Central churches of Central Europe (IX-XI. century) 1984.
- [4] CloudCompare Version 2.6.1 User manual, 2015.
- [5] C.R.Kennedy & Company, 2016: Leica Cyclone Basic User Manual
- [6] CloudCompare Version 2.6.1 User manual, 2015
- [7] John Ashburner & Karl J. Friston, 2004: Rigid Body Registration. London, UK
- [8] John Markoff, 1989: One Man's Fight for Free Software. The New York Times. Business Day, January 11, 1989
- [9] John W. Eaton, David Bateman, Søren Hauberg, Rik Wehbring, 2017: GNU Octave Free Your Numbers. Boston, USA
- [10] Rudolf Staiger, 2003: Terrestrial Laser Scanning Technology, Systems and Applications. 2nd FIG Regional Conference, Marrakech, Morocco

Extension of Nodal Voltage Method with the Thermosensing

György Györök¹, Alexander E. Baklanov², Bertalan Beszedes¹

¹ Óbuda University, Alba Regia Technical Faculty, Budai Str. 45, H-8000 Székesfehérvár
 {gyorok.gyorgy, bertalan.beszedes}@amk.uni-obuda.hu

² D. Serikbayev East Kazakhstan State Technical University
 Instrument Engineering and Technology Automation, Protozanov Str. 69, Ust-Kamenogorsk, Kazakhstan
 ABaklanov@ektu.kz

Abstract—Most of the electronic circuit simulation program contains a components's thermal behavior modeling. Unfortunately in most cases this modeling, or simulation depends on the ambient temperature of suspected electronic devices. This modelings can't use of electronic parts's or or circuit's self warming.

The proposed option, methodology is particularly useful for the simulation of high performance circuits. In this article, we supplement this thermal simulation methods so, that it can not only to take into account the ambient temperature, but also to warm the part itself.

For the demonstration of simulation we use pSpice based MICROCAP software environment.

I. INTRODUCTION

On Fig. (1) is seen a typical application of a high power switching-mode semiconductor circuit. Fig. (2) illustrate a typical input and output signals of suspected (in this case typical theoretical application) circuit [6] [7].

Voltage generator (U_g) drives direct, with 5V, (U_{pp}), gate (G) of F_t transistor (red color of Fig. (2)). If gate-source (S) voltage (U_{GS}) greater then threshold voltage of gate U_{TH} , then transistor saturate and drain (D) voltage (U_D) goes down (blue color of Fig. (2) from power voltage (U_P). Meanwhile, on R_D , I_D current is flowing [15] [3].

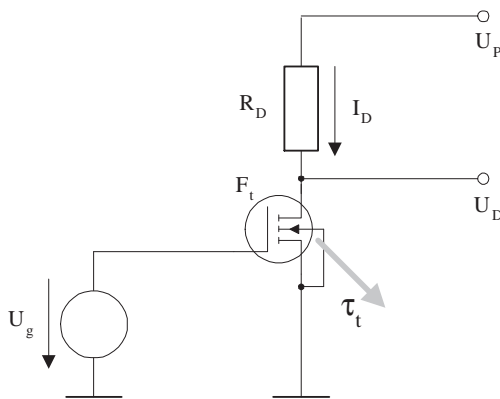


Fig. 1. Typical application of a high power switching-mode semiconductor circuit. ($U_P = 12V$, $R_D = 100\Omega$.) Time function of transistor's dissipation is τ_t .

II. THERMAL BEHAVIOR OF SELF HEATED SEMICONDUCTOR

If the F_t turned ON, it's dissipation depends on channel resistance (R_{DS}), and drain current (I_D). So we get time function of dissipation power according equation (1);

$$p_{F_t}(t) = R_{DS} \cdot I_D(dt)^2. \quad (1)$$

We use differential equation (2) for the description of semiconductor's warming process or heat-work of suspected device;

$$Pdt = cm\tau + S\alpha\tau dt, \quad (2)$$

where c is average specific heat of semiconductor body, m is the mass of heating body, τ is temperature difference of environment and semiconductor device, α is a specific heat transfer factor. We put on the value P constant. Equation (2) solve for P , result is in equation (3)

$$P = cm \frac{d\tau}{dt} + S\alpha\tau. \quad (3)$$

On right side of equation (2) are two components; $S\alpha\tau$ is the quantity which heating environment, and time function of

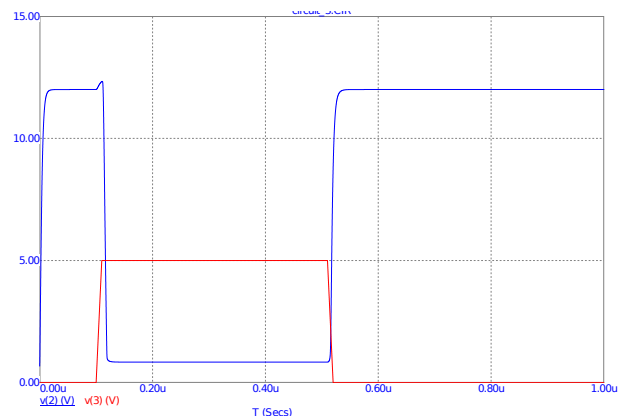


Fig. 2. Typical input and output signals of Fig. (1)'s circuit.

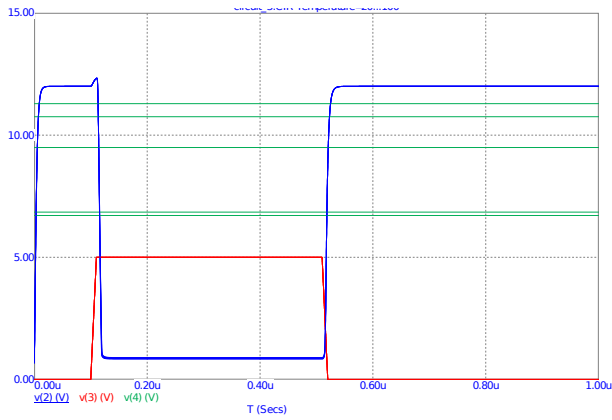


Fig. 3. Output signals of Fig. (1)'s circuit, parameterized in environmental temperature. ΔT is 20°C , green lines.

expression of $cm\frac{dT}{dt}$ gives thermal-work value what heating semiconductor's body itself.

If P constant, from equation 3, we get equation 4;

$$\tau(t) = \tau_0 e^{-\frac{t}{T}} + \tau_m (1 - e^{-\frac{t}{T}}); \quad (4)$$

where $T = \frac{cm}{S\alpha}$ time-constant of heating, $\tau_m = \frac{P}{S\alpha}$ is the heat-balanced overheating, $\tau_0 = \tau(t=0)$ is the initial value of the function.

III. THE USUAL TEMPERATURE SIMULATION

On Fig. (3) is seen a environmental temperature modeling of switching mode high power transistor circuit (Fig. 2). It follow the well used heuristical approach, according which $U_{TH}(\Delta T) \simeq -3mV/^\circ\text{C}$. Exact value of course depends on real type of a semiconductor.

Is seen on Fig. (3) that output characteristics isn't very dependent of ambient temperature, however it is change from $20^\circ\text{C} - 120^\circ\text{C}$.

It can be seen that by conventional simulation we are not really able to model the effect of thermal changes in the semiconductor [14]. This is especially true of the thermal processes on the chip and their effects [11]. If need to take into account the impact of the warming of its assets in other techniques must be used.

IV. EXTEND OF NODAL METHODS

The electronic circuit simulation programs use the nodal potential method. In case of active and passive components, they also use a replacement network. Generally, one branch consists of a voltage generator and branch resistance or impedance (Fig. 4).

We want to determine the voltage of each point (U_p) of the circuit use of equation 5:

$$U_p = R_e \sum_{i=1}^n I_n; \quad (5)$$

where I_n is the branch currents, R_e is parallel sum of branch resistors [5] [2]. By using the Milmann rule we get the equation 6:

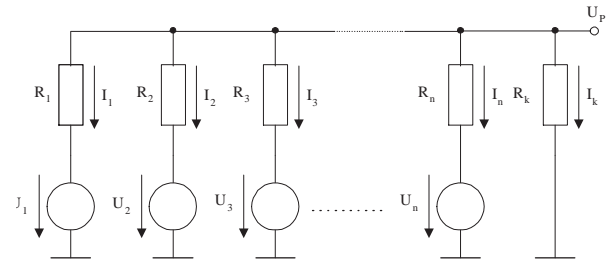


Fig. 4. Interpretation of nodal potential in a general case.

$$U_p = \frac{\sum_{i=1}^n \frac{U_i}{R_i}}{\sum_{i=1}^k \frac{1}{R_i}}; \quad (6)$$

thus, we get a direct correlation between branch voltage generators and branch resistors.

If the temperature change is to be taken into account when determining the value of the nodal potential, the circuit diagram of the figure 4 is supplemented with a temperature dependent branch. The temperature dependent solution is shown in the Fig.5 [4].

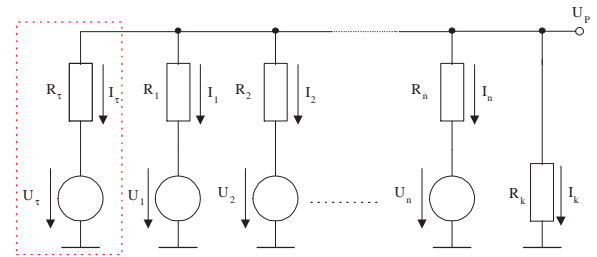


Fig. 5. The temperature dependent solution. Temperature-sensing supplement branch in red dashed box.

Thus, the equation can be supplemented with the equation of the heat sensor branch [9]. So we get equation 7:

$$U_p = R_e \sum_{i=1}^n I_n + R_e I_\tau; \quad (7)$$

where τ is the the temperature-sensing current component. From 7 and 6 we get equation 8:

$$U_p = \frac{\sum_{i=1}^n \frac{U_i}{R_i} + \frac{U_\tau}{R_\tau}}{\sum_{i=1}^k \frac{1}{R_i}}; \quad (8)$$

The element of the branch is the voltage generator and the resistor, each of which can be defined as temperature dependent electronic part [8].

V. CONTROLLED VOLTAGE GENERATOR

By using the equations 4, complete with heuristic results, we can write the voltage generator transfer characteristic equation (9).

$$U_{\tau} = f(\tau(t)dt, h); \quad (9)$$

where; $\tau(t)$ time function of semiconductor heating from equation 4, h is a heuristical constant from earlier experiments, measures.

The physical content of the proposed solution is thermal coupling of a thermal sensor (U_{τ}) with the tested semiconductor whose output voltage is temperature dependent (Fig. V).

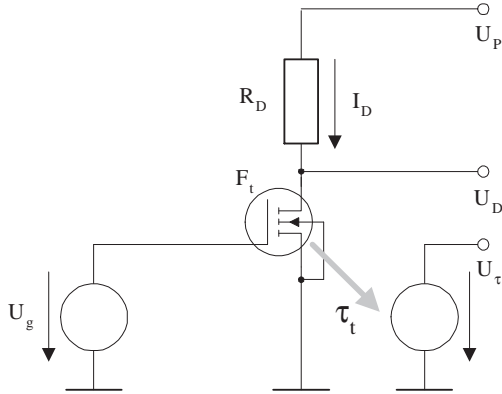


Fig. 6. Thermocouple of a high power transistor (F_t) and a thermos driven voltage generator (U_{τ}).

Such a temperature sensor voltage generator may be a special thermocouple or a semiconductor circuit designed for this purpose (AD22100, Analog Devices [1]). The latter is an forward-mode current generator drive silicon diode, as described in the equation 10:

$$I_D = I_0(T) \left(e^{\frac{U_D}{nU_T}} - 1 \right); \quad (10)$$

where; I_0 is the maximal reverse current, T is temperature, n is the emission coefficient; $1 \leq n < 2$, U_T thermic voltage (equation 11):

$$U_T(T) = \frac{kT}{q}; \quad (11)$$

where; q is the elementary charge, k is Boltzmann-constant. The equations 10, 11 gives the practical used result equation 12

$$\left. \frac{\partial U_D}{\partial T} \right|_{I_D=\text{constant}} = \frac{U_D - U_G - 3U_T}{T}; \quad (12)$$

where; U_G is band gap voltage.

For ΔU_D in practice and from equation 12 is $-1,7mV/^{\circ}C$ largeness change.[12] [13]

As a voltage generator of the Fig. , a current generator driven a forward mode silicon diode is used (Fig.). With its output voltage, connect in serial the U_{GS} voltage of the transistor.

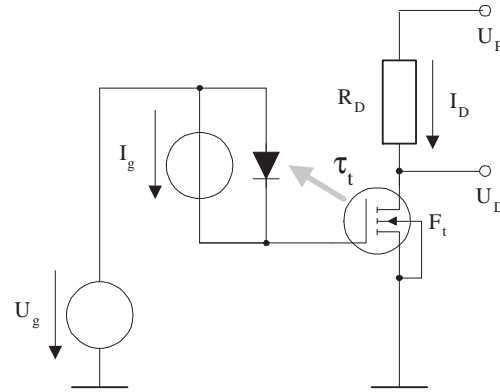


Fig. 7. Thermocouple of a high power transistor (F_t) and a silicon diode.

Thus, we get the dependent output characteristic of semiconductor own heat (Fig.). In the figure, green lines indicate the temperature change.

If we enlarge the important part of the Fig. (8), is seen the different values of the saturation voltage (Fig. 9).

It can be seen that the model works, the output characteristic of the transistor is temperature dependent.

VI. USE OF HEAT SENSITIVE RESISTOR

If a constant-voltage generator is used, we assume that resistor is temperature dependent. In this case, the resistor (R_T) itself is thermal coupled with the power transistor (Fig. 10). In the case of thermistors, we are approaching the change of resistance with the usual formulas, equation 13:

$$R(T) = R_{\infty} e^{\frac{B}{T}}; \quad (13)$$

where B is thermal sensing index in equation 14

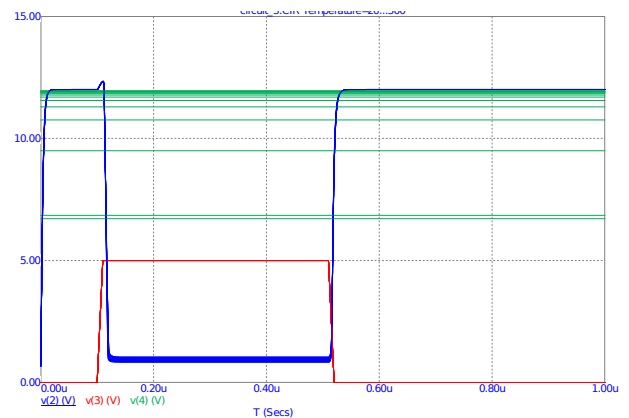


Fig. 8. Output characteristics with modified U_{TH} voltage parameters.

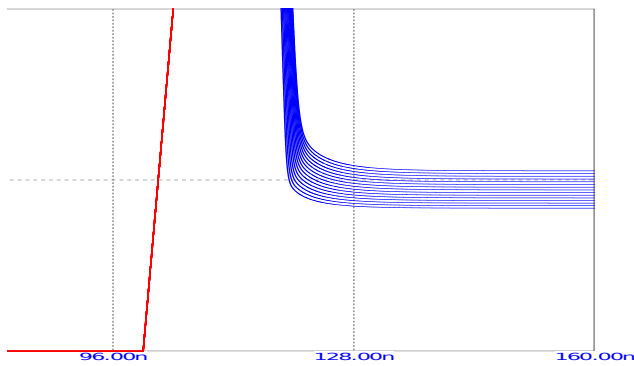


Fig. 9. The relevant enlarged part of turn ON curve of Fig. (8).

$$B = \frac{T_2 T_1}{T_2 - T_1} \ln \frac{R_1}{R_2}; \quad (14)$$

where R_∞ is in equation 15

$$R_\infty = R_1 e^{\frac{B}{T_1}}; \quad (15)$$

where $R_1 = R_{T=20^\circ C}$ and $R_2 = R_{T=100^\circ C}$ [10].

We can proceed similarly to the use of thermistor as in Fig. by diode. Of course, we can choose another mode to change the transistor driven by changing the resistance of the thermistor. In this case we need to build different of U_g and U_τ and this driven to transistor's U_{GS} voltage; $U_{GS} = U_g - U_\tau$.

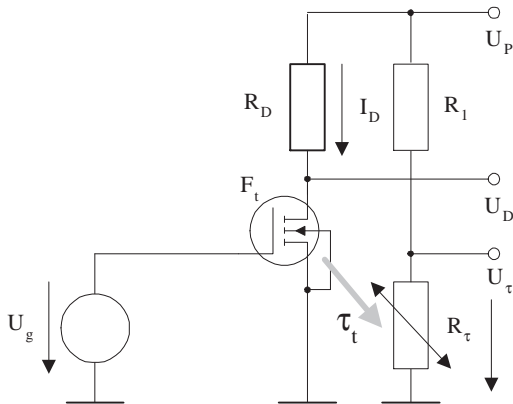


Fig. 10. Thermistor as a heating sensor.

VII. CONCLUSIONS

Present article is part of a lengthens future work in which we intend to supervise high-reliability electronic circuits with a microcontroller. To do this, we perform a modeling process in which a circuit simulation method compares the test circuit to some of its voltages, theoretically calculated. To do this, thermal modeling is very useful, and through some simple examples we can prove that the procedure works, it can be used.

REFERENCES

- [1] Analog Devices. AD22100 voltage output temperature sensor with signal conditioning. <http://www.datasheetcatalog.com/datasheets.pdf/AD/2/2/AD22100.shtml>, 6(1):151–160, 2009.
- [2] Gy. Györök, M. Makó, J. Lakner. Combinatorics at electronic circuit realization in FPAA. *Acta Polytechnica Hungarica, Journal of Applied Sciences*, 6(1):151–160, 2009.
- [3] Györök György. *Programozható analóg áramkörök mikrovezérlő környezetben*. Óbudai Egyetem, ISBN 978 615 5018 97 8, Budapest, 2013.
- [4] Györök György. *Számítógép perifériák I*. Óbudai Egyetem, OE AREK 8003 ISBN 978 615 5018 57 2, Budapest, 2013.
- [5] Gy. Györök. Self configuration analog circuit by FPAA. *Proc. 4th Slovakien–Hungarian Joint Symposium on Applied Machine Intelligence (SAMJ2006)*, pages 34–37, January 2006.
- [6] Gy. Györök. A-class amplifier with FPAA as a predictive supply voltage control. *Proc. 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI2008)*, pages 361–368, November 2008.
- [7] Gy. Györök. Crossbar network for automatic analog circuit synthesis. *Proceedings (Liberios Vokoros, Ladislav Hluch, Jnos Fodor szerk.) of the IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI 2014)*. IEEE Computational Intelligence Society, Budapest: IEEE Hungary Section, ISBN:978-1-4799-3441-6, pages 263–267, January 2014.
- [8] J. Kopják. Dynamic analysis of distributed control network based on event driven software gates. *IEEE 11th International Symposium on Intelligent Systems and Informatics, Subotica, Serbia*, ISBN: 978-1-4673-4751-8:p. 293–297, 2013.
- [9] J. Kopják and J. Kovács. Implementation of event driven software gates for combinational logic networks. *IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, Subotica, Serbia*, ISBN: 978-1-4673-4751-8:p. 299–304, 2012.
- [10] Power Electronics News. <http://www.electronics-tutorials.ws/io/thermistors.html>. pages 263–267, 2014.
- [11] T. Orosz. Analysis of sap development tools and methods. *15th IEEE International Conference on Intelligent Engineering Systems (INES)*, pages pp. 439–443, 2011.
- [12] A. Pilat and J. Klocek. Programmable analog hard real-time controller [programowalny sterownik analogowy]. *Przegląd Elektrotechniczny*, 89(3 A):38–46, 2013. cited By (since 1996) 0.
- [13] Adam Pilat. Control toolbox for industrial programmable analog controllerembedding state feedback controller. pages 1–4, 2012.
- [14] A. Selmeci and T. Orosz. Usage of soa and bpm changes the roles and the way of thinking in development. *IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pages pp. 265–271, 2012.
- [15] Poppe A. Székely V. *Áramkör Szimuláció PC-n*, Budapest, Computer Books, 1996. ISBN 963 618 080 6.