Vol. 4. Issue 1, 2018 ISSN: 2453-7314

Central European Researchers Journal

CERES

A Word of Welcome from the Editors

Dear Colleagues, Readers and Authors,

We are pleased to present you next issue of the "Central European Researchers Journal (CERES)". The journal presents articles that deal with information technologies, broadly construed. It features papers that crosses the boundaries of a variety of disciplines, including different applications. The main purpose of the journal is not to develop a particular discipline but to advance young researchers investigations of the multifaceted reality with the application of information technologies and for the decision of problems of the information technologies.

This issue is prepared according to traditions that were formed under the project CERES (544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES): to create and support the platform for the cooperation of academic community and first of all of young researchers from EU and partner countries. The issue consists of articles of authors from Belarus, Slovakia and Ukraine. All published paper are reviewed by two or more reviewers. The publications are open access and free of charges.

With best wishes

Prof. Vyacheslav Kharchenko
Prof. Elena Zaitseva

Central European Researchers Journal. Volume 4. Issue 1
Editor-in-chief: Kharchenko Vyacheslav, Zaitseva Elena

Addres of the editorial office: Central European Researchers Journal - editorial, Faculty of Management Science and Informatics, University of Zilina, Univerzinta 8215/1, 01026, Zilina, Slovakia, editorial@ceres-journal.eu
Each paper was reviewed by reviewers.
Publisher: JMTM, s.r.o., Sad SNP 8, 010 01, Zilina, Slovakia, publisher@jmtm.sk
Published biannually
ISSN: 2453-7314
July 2018

Contents

Determining the Influence of the External Factors on the Firmware Response Time

Dmytro V. Fedasyuk, Vasyl I. Gavrysh, Tetyana A. Marusenkova, Ratybor S. Chopey

***Abstract*—**The work deals with the problem of changes in the response time of hardware components of a real-time embedded system over time and under the influence of the environment. Simple models for predicting the change of the response time under influence of the supply voltage, the ambient temperature and the operation time of an electronic unit being evaluated have been proposed. The models sustained approbation with experiments.

***Keywords*—**firmware execution time, hardware components, real-time embedded systems

I. Introduction

In hard real-time systems, each time-critical activity should meet its deadline. However, any firmware execution time depends not only on the microcontroller itself but also on peripheral devices connected to it, moreover, the latter can be inclined to more or less uncertainty in their response, depending on their type, model, the time of being in use, and the conditions of use. It's commonly known that each electronic unit is getting slower with age. Thus, any firmware's execution time depends not only on the computational resources of the embedded system being inspected but also on peripheral devices each of which is able to perform a set of operations during some random time period residing, however, within a known interval. Moreover, the ambient temperature influences a device, thus the latter might work slower in hot areas than in cold or moderate ones.

In order to evaluate the firmware's execution time, they use the following metrics [1]: worst-case execution time (WCET), best-case execution time (BCET) and average-case execution time, (ACET). The latter resides within the interval [BCET;WCET] and depends on the distribution of the program execution time. Evaluation of the firmware's execution time is typically performed *before* a system gets being in use and its results are considered invariable. However, the firmware's response time might change over time and in the case when the system's environment changes. The firmware's response time is unlikely to get better, thus a system previously considered useable might become no longer applicable some time moment during its operation, and this fact should be taken into account.

A range of methods has been developed for estimation of firmware execution time. They all fall into two large groups: static methods and dynamic ones. These methods are based either on the measurement of the actual execution time of firmware in a real embedded system [2], [3] or on predicting the range in which the firmware execution time resides [4]. References [5] – [7] present a detailed overview of methods and software tools intended for analysis of the execution time of a firmware.

None of the existing methods takes into consideration the influence of changes in external or internal physical quantities (supply voltage, ambient temperature, aging of the hardware components comprising an embedded system under evaluation and the fact that hardware gets worn-out over time). The work is aimed at the development of a model for evaluation of firmware execution time, that takes into account changes in the response time of electronic units

D. V. Fedasyuk, Lviv Polytechnic National University, Lviv, Ukraine (e-mail: fedasyuk@gmail.com).
V. I. Gavrysh, Lviv Polytechnic National University, Lviv, Ukraine (e-mail: ikni.pz@gmail.com).
T. A. Marusenkova, Lviv Polytechnic National University, Lviv, Ukraine (e-mail: tetyana.marus@gmail.com).
R. S. Chopey, Lviv Polytechnic National University, Lviv, Ukraine (e-mail: chopey.ratybor@gmail.com).

under influence of changes in the supply voltage, ambient temperature, and aging.

II. Problem Statement

The text of any firmware that is comprised of a set of instructions, might be represented by a control flow graph (Fig. 1).

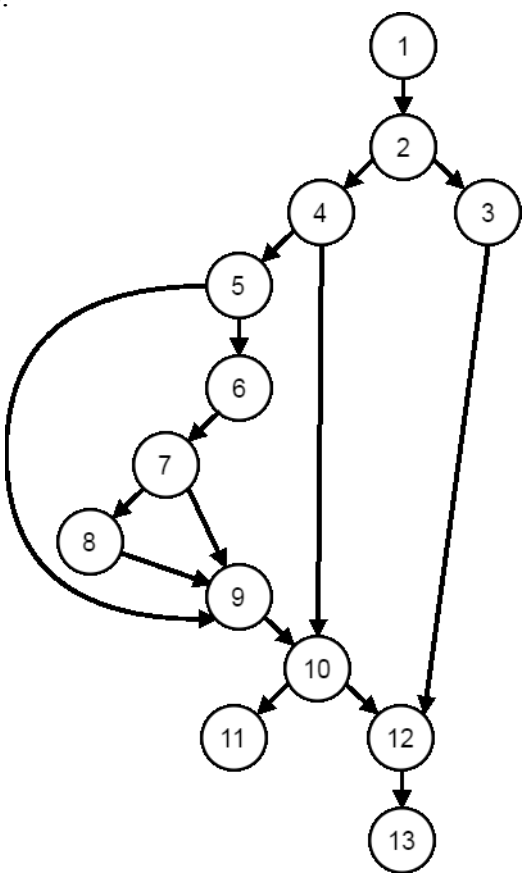


Fig. 1. An example of a control flow graph representing the code of a firmware function

The vertices of a control flow graph represent operators or functions that are characterized by some execution time. The edges show connections between different operators/functions. The whole execution time of a firmware code branch can be calculated by the following formula:

$$\tau_b = \sum_{i=0}^{n} \tau_i \quad (1)$$

where τ_i is the execution time of a certain vertex in the graph, n is the amount of vertices in a single path in the firmware being analyzed.

Making allowances for the fact that the execution time of any firmware depends not only on the computational resources of the underlying microprocessor but also on the response time of the peripheral devices included into the embedded system under investigation, one can describe the resulting execution time as follows:

$$\tau_b = \sum_{j=0}^{k} \tau_c + \sum_{i=0}^{n} \tau_v\,(param) \quad (2)$$

where τ_c is the execution time of a graph vertex that depends on the computation capabilities of the processor only; k is the amount of computational instructions with the constant execution time; $\tau_v(param)$ is the execution time of a graph vertex which depends on the response time of a peripheral device and may vary under influence of some internal or external physical quantity, n is the amount of computational operations with a variable execution time.

III. Experimental Data

In order to verify our assumption that the influence of the temperature, supply voltage, and age on the response time of an electronic unit is of any importance and is worth being taken

into consideration, we started our investigation with a set of experiments with a thermocouple ADC ADS1118, and seeking for data on the dependence of the response time on aging of hardware components. The same data will be used for verification of analytical dependencies derived in this work.

Table I shows the dependency of the ADS1118's response time on the supply voltage. The nominal response time (in accordance with the official datasheet [8]) is 0.0078125 s. It can be easily seen that the discrepancy between the nominal response time and the actual one is up to 1.2% (the minus sign shows that the actual value is greater than the nominal one).

The data summarized in Table II do not represent the authors' experimental results; they were taken from [9]. One can notice that 3 months of operation reduce the response time by 10% whereas three years under operation worsen the response time by 18%.

Table III presents measurements of ADS1118's response time at different temperatures for two cases: 1) the supply voltage is fixed to 2V; 2) the supply voltage is fixed to 3.3 – 5V taken from [8]. The variance of the response time comes up to 2.4% in our specific case. The percentage itself might differ depending on electronic units being considered.

However, the main point remains the same – the unit's response time does depend both on temperature and supply voltage, as proved by experiments conducted by the authors and provided by the manufacturer.

When setting a measurement mode for ADC ADS1118, one of the parameters to be set is the number of conversions per second (and, consequently, the duration of any single conversion, or the response time).

TABLE I
THE DEPENDENCE OF THE ADS1118'S RESPONSE TIME ON THE SUPPLY VOLTAGE

Supply voltage, V	Response time, s	The difference between the nominal and actual values, %
2	0.00771875	-1.2
2.1	0.00771875	-1.2
2.2	0.007726563	-1.1
2.3	0.007734375	-1
2.4	0.007734375	-1
2.5	0.007742188	-0.9
2.6	0.007757813	-0.7
2.7	0.007773438	-0.5
2.8	0.007789063	-0.3
2.9	0.007804688	-0.1
3	0.0078125	0
3.1	0.007828125	0.2
3.2	0.007835938	0.3
3.3	0.00784375	0.4
3.4	0.007844531	0.41
3.5	0.007844531	0.41
3.6	0.007847656	0.45
4	0.007849219	0.47
4.5	0.00785	0.48
5	0.007851563	0.5

TABLE II
THE DEPENDENCE OF THE ADS1118'S RESPONSE TIME ON THE TIME OF BEING IN USE

Operation time, years	Response time, s	The difference between the nominal and actual values, %
0.25	0.00859375	10
0.6	0.00859375	10
1.7	0.00859375	10
1.9	0.008984375	15
2.2	0.008984375	15
3	0.00921875	18

TABLE III
THE DEPENDENCE OF THE ADS1118'S RESPONSE TIME ON THE AMBIENT TEMPERATURE AT THE SUPPLY VOLTAGE 2 V AND 3,3/5 V

Supply voltage, V	Response time, s (2 V)	The difference between the nominal and actual values, %	Response time, s (3,3/5 V)	The difference between the nominal and actual values, %
-60	0.007625	-2.4	0.00789063	1
-40	0.007648438	-2.1	0.00787500	0.7
-20	0.007671875	-1.8	0.00786719	0.6
0	0.007695313	-1.5	0.00785156	0.5
20	0.007710938	-1.3	0.00784375	0.4
40	0.007734375	-1	0.00784375	0.3
60	0.007746094	-0.85	0.00783594	0.3
80	0.007773438	-0.5	0.00783594	0.25
100	0.007789063	-0.3	0.00782813	0.25
120	0.007804688	-0.1	0.00782813	0.3
140	0.007820313	0.1	0.00783594	0.3

In order to have the possibility to predict the difference between the nominal response time and actual one for any value of the supply voltage, ambient temperature or time of being in use, one needs some analytical dependencies.

IV. DEVELOPING A MODEL

Having analyzed the above-presented data, one can assume that the sought-for analytic dependencies might be sufficiently well approximated by linear dependencies:

$$\tau = \alpha_n + \beta_n x, \tag{3}$$

where x – is the value of a physical quantity that influences the response time of an electronic unit (temperature, supply voltage, or aging period), α and β are the coefficients to be found.

In order to find the most suitable values for the coefficients, we have used LMS and derived the following formulas:

$$\alpha = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \tau_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \tau_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \quad \beta = \frac{n \sum_{i=1}^{n} x_i \tau_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \tau_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \tag{4}$$

where τ_i – is the corresponding measured value of the response time (or taken from the technical reports).

Using formulas (4), we've calculated:

1) The coefficients that describe the influence of the supply voltage on the response time:

$$\alpha_1 = 7{,}622 \cdot 10^{-3};\ \beta_1 = 5{,}674 \cdot 10^{-5}; \tag{5}$$

2) The coefficients describing the influence of electronic units' aging on their response time:

$$\alpha_2 = 8{,}455 \cdot 10^{-3};\ \beta_2 = 2{,}317 \cdot 10^{-4}; \tag{6}$$

3) The coefficients describing the influence of the ambient temperature on the response time at the supply voltage 2V

$$\alpha_3 = 7{,}69 \cdot 10^{-3};\quad \beta_3 = 9{,}747 \cdot 10^{-7}; \tag{7}$$

4) The coefficients describing the influence of the ambient temperature on the response time at the supply voltage 3.3 – 5 V

$$\alpha_4 = 7.85994 \cdot 10^{-3};\ \beta_4 = -2.80527^{-7}; \tag{8}$$

Having substituted the calculated coefficients (5-8) in (3), we obtain the response time of a peripheral device influenced by each considered physical quantity. The result should be related to (2).

V. APPROBATION

In order to verify the obtained formulas, the following several steps were taken. First, we compared two plots for the dependence of τ on the supply voltage: one based on experimental results, another – on analytical dependencies (just for preliminary and very rough estimation).

Then, at the second step, we artificially reduced the number of experiments (n) from n = 20 to n = 15 and n = 10, i.e., we intentionally did not use part of experimental data for calculating the coefficients. Then we verified the differences between the values of τ calculated analytically, using the above-presented formulas, and the corresponding values of τ obtained experimentally. As expected, we noticed that the inaccuracy grows with reduction of data used for calculating of the coefficients. Finally, we performed several additional measurements (the results were not included in the calculation of the coefficients) to check how they fit the analytical curve. This approach is represented in Fig. 2 – Fig. 5. The same procedure was repeated for two remaining dependencies. The greatest calculation error for the dependency on the supply voltage is $5.4e^{-5}$, on the time of being in use – $2.56e^{-4}$, on the ambient temperature at the supply voltage 2V – $6.57e^{-6}$, on the ambient temperature at the supply voltage 3.3 – 5V – $1.52e^{-5}$, and this fact proves the applicability of the proposed model.

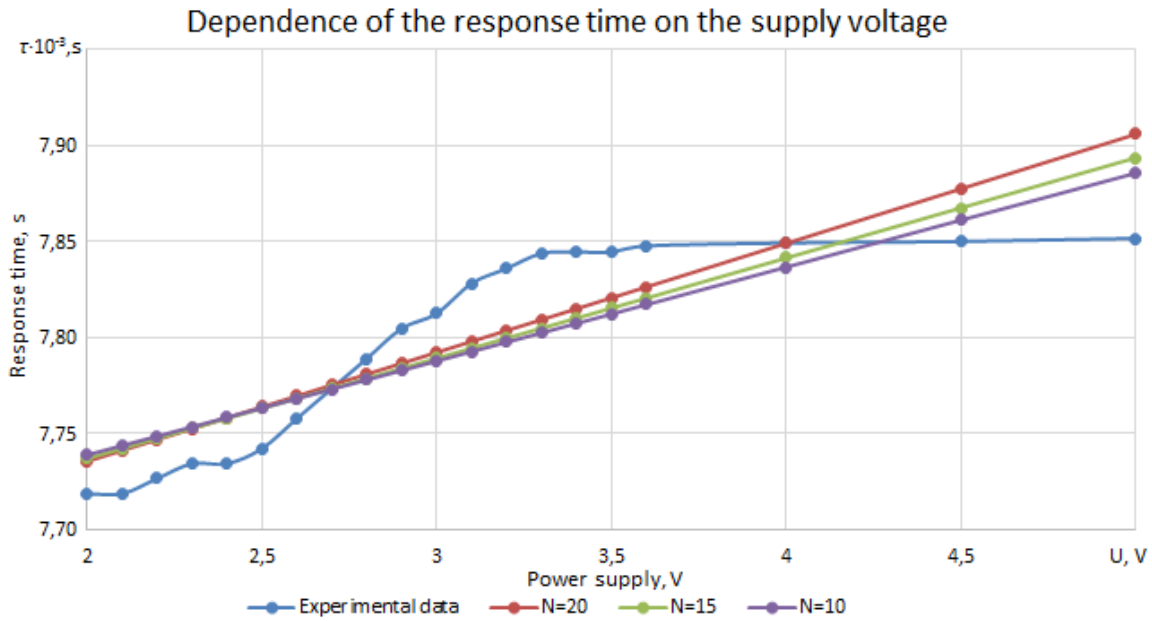


Fig. 2. The experimental and calculated dependencies of the response time on the supply voltage

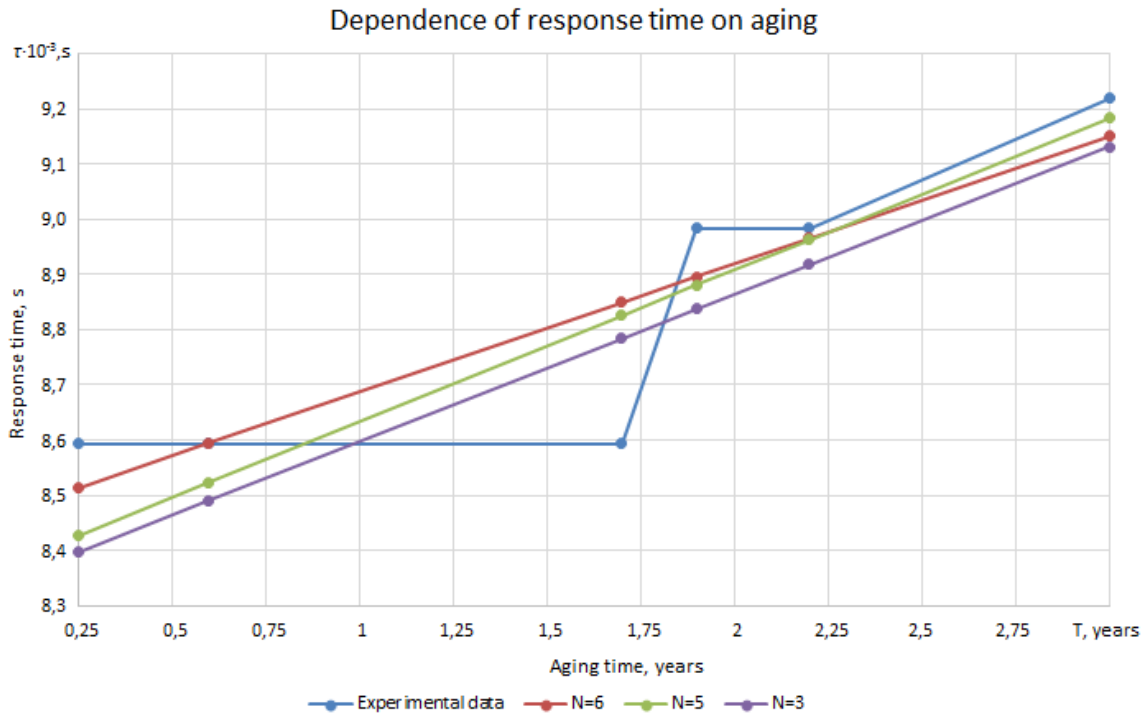


Fig. 3. The experimental and calculated dependencies of the response time on the time of being in use

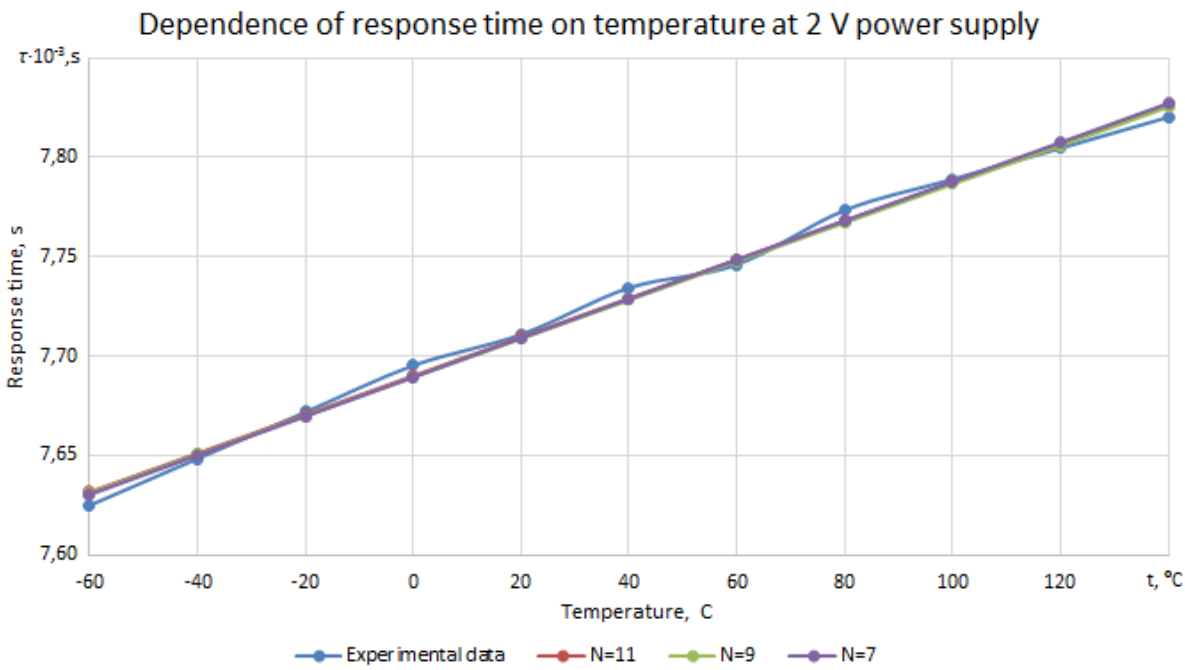


Fig. 4. The experimental and calculated dependencies of the response time on the temperature at the supply voltage 2V

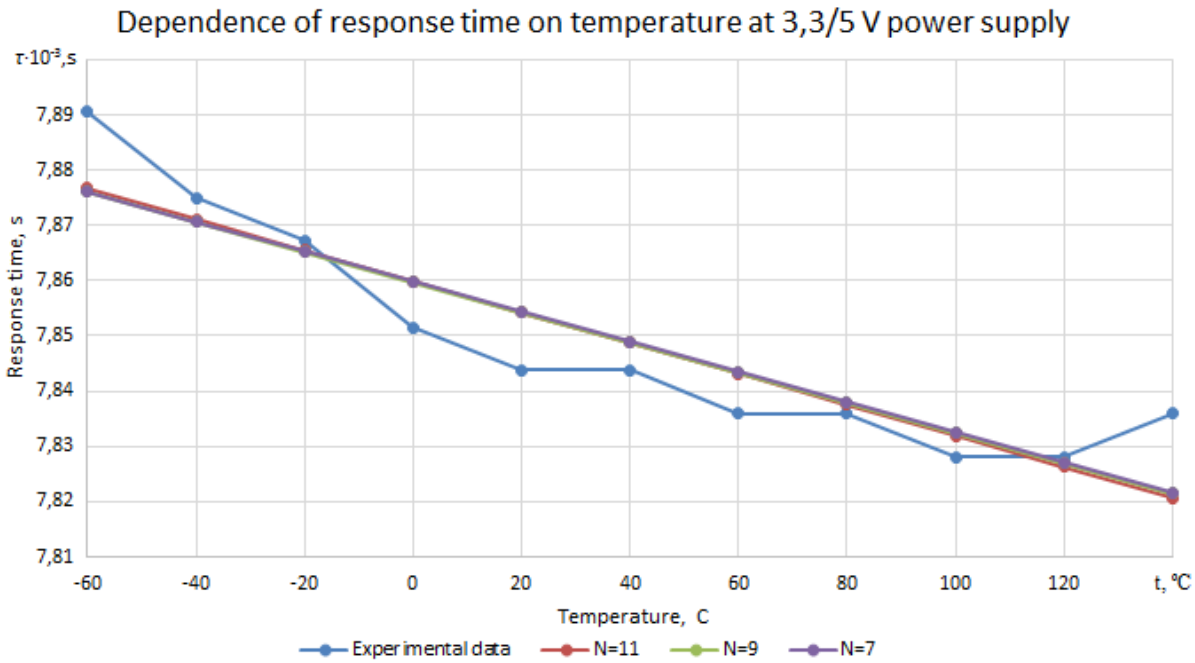


Fig. 5. The experimental and calculated dependencies of the response time on the temperature at the supply voltage 3.3 – 5 V

The models provided for calculating the coefficients are universal, i.e. they can be applied for any electronic units showing similar dependencies of the response time on the supply voltage, the time being in use and the ambient temperature. However, the coefficients themselves are unique for a specific electronic unit model and, likely, for each specimen as well. It might be, however, that the characteristics of different specimens of the same electronic unit vary within some small and acceptable range. Thus, to be able to apply the proposed model, one has to perform a set of experiments in order to obtain the results like those presented in Tables 1, 2 and 3 first and then to calculate the coefficients upon the conducted measurements.

It's worthwhile to store the coefficients in some reliable external memory chip of an

embedded system in order to allow the latter to evaluate the reliability and applicability of the response time of its components. I.e., knowing its own properties and the properties of its components, an embedded system might be able to warn its operator that some operations slowed down and are no longer fast enough for normal operation.

VI. Conclusion

We've obtained three independent analytic dependencies of the difference between the nominal response time of an electronic unit and its actual value on the supply voltage, the ambient temperature and the time of being in use. The proposed dependencies do not allow us to find out the response time using all the three parameters (supply voltage, temperature, and age) simultaneously, i.e. each dependency shows the contribution of the only quantity. We are planning further investigation aimed at forming a unique function of three variables that would allow us to calculate the resulting response time.

The practical use of the proposed model is as follows. We use the derived formulas (3-4) independently and select the worst result (maximum $\Delta\tau$). Two other quantities can only worsen this result but never improve it. If the worst result proves that the system is no longer applicable, its operator should take some predefined actions. If not, there is not enough information to make any conclusions.

On the other hand, if we evaluate the increases in the response time contributed by each physical parameter individually, then sum them up and the sum does not make the response time go out of the expected range, we can conclude that the response time still fits the established norm. Here we rely on the fact that even if age, supply voltage and temperature do mutually impact each other, this influence has been already taken into account when drawing the dependencies between the response time and each of these parameters individually.

Acknowledgment

The authors thank the stuff of Dinamica Generale S.p.A. for their consistent support and sharing experience.

References

[1] Embedded Software: Know It All / [J. Labrosse, J. Ganssle, R. Oshana, C. Walls, K. Curtis, J. Andrews, D. Katz, R. Gentile, K. Hyder, B. Perrin]. – Burlington: Newnes, 2008. – 745 p. ISBN-13: 978-0750685832

[2] F. Wartel, L. Kosmidis, C. Lo, B. Triquet, E. Quinones, J. Abella, A. Gogonel, A. Baldovin, E. Mezzetti, L. Cucu et al., "Measurement-based probabilistic timing analysis: Lessons from an integrated-modular avionics case study," in Proceedings of the 8th IEEE International Symposium on Industrial Embedded Systems. IEEE, 2013, pp. 241–248.

[3] B. Lesage, D. Griffin, F. Soboczenski, I. Bate, and R. I. Davis, "A framework for the evaluation of measurement-based timing analyses," Proceedings of the 23rd International Conference on Real-Time Networks and Systems, 2015.

[4] Supporting timing analysis by automatic bounding of loop iterations / [C. Healy, M. Sjodin, V. Rustagi et al.] // Real-Time Systems. – 2000. – Vol. 18, № 2-3. – P. 129–156. DOI: 10.1023/A:1008189014032

[5] R. Wilhelm, J. Engblom, A. Ermedahl, N. Holsti, S. Thesing, D. Whalley, G. Bernat, C. Ferdinand, R. Heckmann, T. Mitra et al., "The worst-case execution-time problem — overview of methods and survey of tools," ACM Transactions on Embedded Computing Systems (TECS), vol. 7, no. 3, p. 36, 2008.

[6] Andreas Ermedahl and Jakob Engblom. Execution time analysis for embedded real-time systems. Handbook of Real-Time Embedded Systems, 2007.

[7] L. Tan, "The Worst Case Execution Time Tool Challenge 2006: The External Test", Second International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (Isola 2006), 2006. DOI: 10.1109/ISoLA.2006.73

[8] ADS1118 Ultrasmall, Low-Power, SPI™-Compatible, 16-Bit Analog-to-Digital Converter with Internal Reference and Temperature Sensor. Texas Instruments, 2013.

[9] U.S. Nuclear Regulatory Commission, "Effect of Aging on Response Time of Nuclear Plant Pressure Sensors", Washington DC, 1989.

Reflections on Marketing Strategy and Digital Marketing Infrastructure along the Belt and Road Project

Anastasia Yurkovskaya, Irina Prishchepa, Semen Losmakov, Uladzimir Parkhimenka

***Abstract*—** The Belt and Road initiative has been analyzed within the framework of marketing channels management. Using the concept of marketing channels flows, global value chains, and push/pull strategies, authors have shown the necessity of incorporation of marketing methodology in the Belt and Road strategy. The need and directions of digital marketing tools implementation on different levels of the Belt and Road initiative has been described. A case of special interest (China-Belarus Industrial Park) has been analyzed from the marketing strategy perspective.

***Keywords*—**Belt and road initiative, digital marketing infrastructure, global value chains, marketing strategy.

I. Marketing Perspective of the Belt and Road

The Silk Road Economic Belt and the 21st-century Maritime Silk Road (also known as The Belt and Road) is a development strategy and framework, proposed by Chinese President Xi Jinping that focuses on connectivity and cooperation among countries primarily between China and the rest of Eurasia. It consists of two main components, the land-based "Silk Road Economic Belt" and oceangoing "Maritime Silk Road" [1].

The routes cover more than 60 countries and regions from Asia to Europe via Southeast Asia, South Asia, Central Asia, West Asia and the Middle East, currently accounting for some 30 per cent of global GDP and more than 35 per cent of the world's merchandise trade. By 2050, the Belt and Road region aims to contribute 80 per cent of global GDP growth, and advance three billion more people into the middle class [2].

The five major goals of the Belt and Road initiative are policy co-ordination, facilities connectivity, unimpeded trade, financial integration, and people-to-people bonds [3].

The core of the Belt and Road initiative is to build a network of infrastructure projects (“roads and ports”) across Eurasia to encourage trade.

Geographically and logistically (see Figure 1), the Belt and Road initiative bases on to-be constructed six international economic co-operation corridors, and the maritime silk road [4]:

1. New Eurasian Land Bridge – from Western China to Western Russia;
2. China–Mongolia–Russia Corridor – from Northern China to Eastern Russia;
3. China–Central Asia–West Asia Corridor – from Western China to Turkey;
4. China–Indochina Peninsula Corridor – from Southern China to Singapore;
5. Bangladesh-China-India-Myanmar Corridor – from Southern China to India;
6. China–Pakistan Corridor – from South-Western China to Pakistan;
7. Maritime Silk Road – from the Chinese Coast through Singapore to the Mediterranean.

From the marketing point of view these corridors should be understand not only in technical way as a transportation infrastructure, but first of all as marketing channels from China to the rest of Eurasia and Africa and vice versa.

Anastasia Yurkovskaya, Belarusian State University of Informatics & Radioelectronics, Minsk, Belarus (e-mail: ap.yurkovskaya@gmail.com).
Irina Prishchepa, Belarusian State University of Informatics & Radioelectronics, Minsk, Belarus (e-mail: irprishchepa1997@gmail.com).
Semen Losmakov, Belarusian State University of Informatics & Radioelectronics, Minsk, Belarus (e-mail: S.losmakov@gmail.com).
Uladzimir Parkhimenka, Belarusian State University of Informatics & Radioelectronics, Minsk, Belarus (e-mail: parkhimenko@bsuir.by).

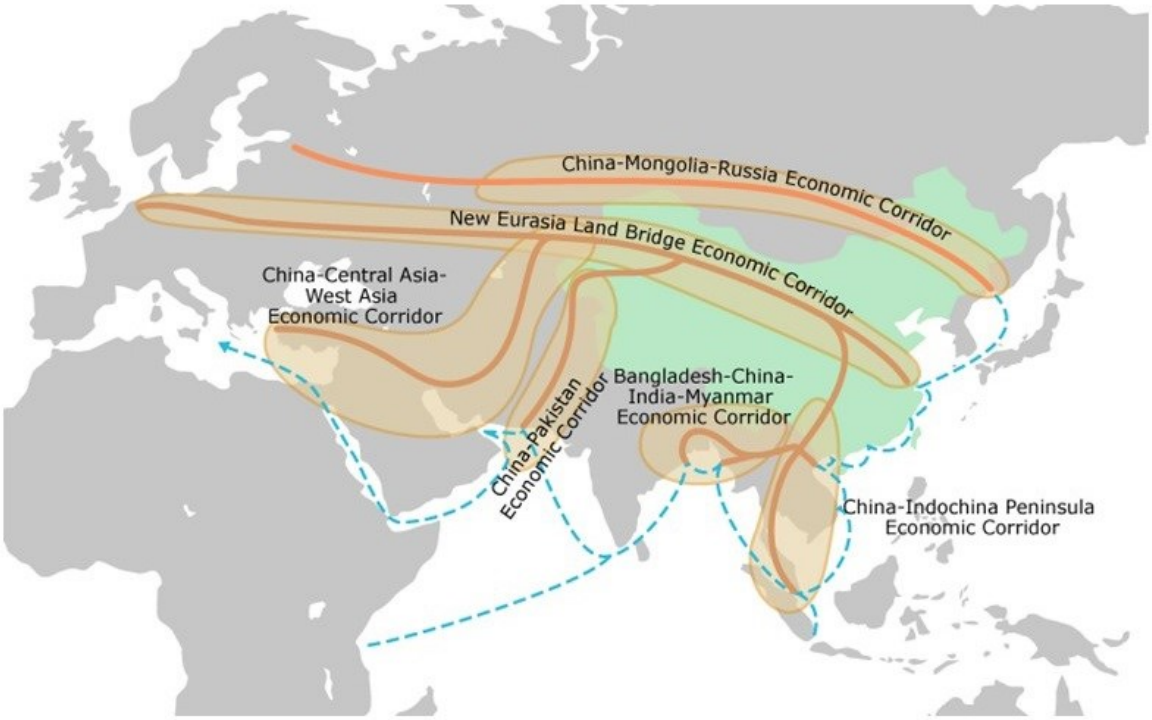


Fig. 1 The Belt and Road's six corridors and Maritime Silk Road on the map [5]

A marketing channel seen from a company's perspective may be defined as "the external contractual organization that management operates to achieve its distribution objectives" [6]. Another approach is to define a marketing channel as a set of people, organizations, and activities necessary to transfer the ownership of goods from the point of production to the point of consumption.

Thus, the main emphasis should be moved from the logistic question (how to transport, cheaply and on time) to the final goal of any marketing channel – a purchase, and through a purchase to customer satisfaction, repeat purchases, customer retention and customer loyalty, and finally to long-term win-win relationship and partnership.

In order to show the possible role of marketing "along" the Belt and Road and to find perspective directions of marketing methodology implementation, we use three classical marketing concepts: marketing channels flows concept, global value chain concept, and push/pull strategies concept.

II. Marketing Channels Flows Concept and the Belt and Road

Please Within a marketing channel, one always can find several flows in addition to the transportation (physical) flow, which can be seen as a natural and necessary one, but nowadays not the most important for a market success.

In the world-known textbook, Kotler and Keller distinguish five major flows in any marketing channel [7]:

1. Physical flow (movement of goods and services)
2. Title flow (movement of property rights)
3. Payment flow (movement of money)
4. Information flow (movement of data)
5. Promotion flow (movement of promotion tools)

Some authors speak about even more number of flows, for example, about "the eight generic

channel flows": 1. Physical possession, 2. Ownership, 3. Promotion, 4. Negotiation, 5. Financing, 6. Risking, 7. Ordering, 8. Payment [8].

The concept of channel flows provides the basis for distinguishing between general channel strategy and transportation (logistics) management. General channel strategy and management involve planning for and managing all of the flows, whereas logistics is concerned almost exclusively with the management of the product (physical) flow [6].

Therefore, a marketing channel should be viewed as a set of flows beyond mere transportation, with the dominant role of promotional flow in it.

The concept applied to the Belt and Road supposes multi-country, cross-border, cross-regional, cross-cultural character (see Fig. 2).

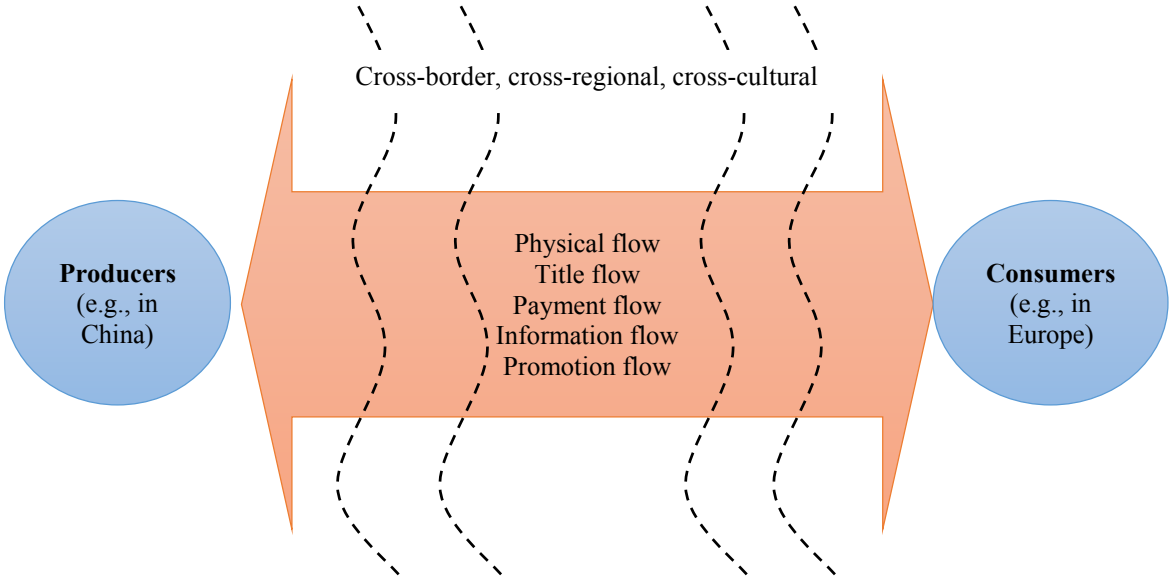


Fig. 2 Marketing channel flows model along the Belt and Road

This implies a necessity of unifying not only transportation infrastructure, but also banking, custom, information and marketing infrastructure, systems, strategies, and activities. In order to make a channel work you should manage all flows through different countries, different legal, custom, banking and all other "regimes", as well as cope with substantial difference in language and consumer behavior and psychology.

III. Push & Pull Strategies Concept and the Belt and Road

In order to "generate" a promotional flow in the channel a company participating in the Belt and Road has used marketing in its own way. The good starting point of thinking about a possible way and form of marketing implementation is the push & pull strategies concept.

Push & pull strategies are promotional (marketing) strategies used to get the product to its target market. A push strategy places the product in front of the customer by one of many types of advertisement to make sure the consumer is aware of the existence of the product. This type of strategy works for low value items and impulse buy items. A pull strategy stimulates demand and motivates target customers to actively look for a specific product. It is aimed primarily at the end users. A strong and well-known brand is needed to ensure the success of a pull strategy.

Applying this definition to the Belt and Road (see Fig. 3), we should speak about two principal ways of marketing (with many hybrid forms between):

1. producers (e.g., Chinese ones) could be oriented to strong brand creation and its direct promotion (e.g., via internet) across all borders to the end consumers (e.g., European), thus

realizing the pull strategy;

2. producers (e.g., Chinese ones) are less active in marketing to the end consumers (e.g., European), and different intermediaries (possibly in the country where a target market is) actively push a product (possibly under their own brand).

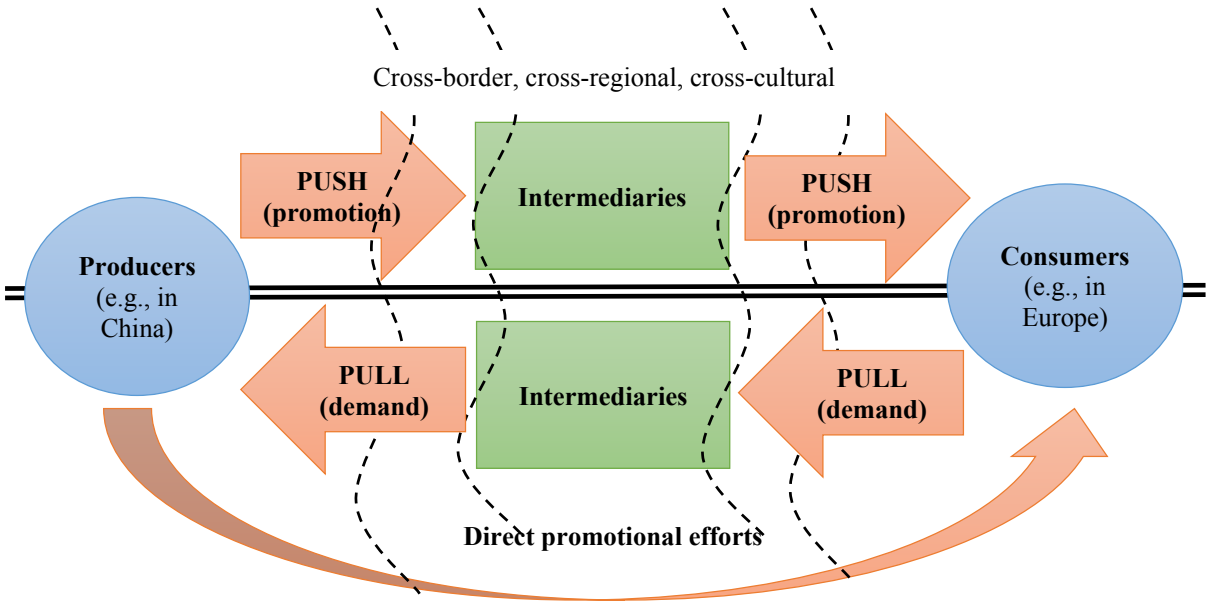


Fig. 3 Push & pull strategies model along the Belt and Road

Push/pull strategies model along the Belt and Road necessarily incorporates cross-country, cross-border, cross-regional and cross-cultural character which makes promotional efforts not so easy to implement.

As a basic rational strategy, it could be recommended to use push-marketing when (a) a product is more complicated, (b) there is a need of additional services for a client, (c) difference in language and consumer behavior / psychology is huge. Pull-marketing seems to be rational in the case of a producer of B2C-market commodities, but it could not be the clear-cut solution.

IV. Global Value Chain Concept and the Belt and Road

Another concept that should be taken into the account is the global value chain approach.

The value chain describes the full range of activities that firms and workers do to bring a product or service from its conception to its end use and beyond. This includes activities such as design, production, marketing, distribution and support to the final consumer. A global value chain is divided among multiple firms and geographic spaces [9].

All said above is fully applicable to the Belt and Road global project, which should be considered, after the construction is finished, as a set of multiple global value chains across Eurasia (see Fig. 4).

In the context of this research, the most important issue is the need of marketing, sales and services to the end consumers in every global value chain to be functional and profitable.

In addition, there is a strategic marketing problem for any participant of such global value chains along the Belt and Road to find its “place” within a specific chain thus using its competitive advantages, adding value to the final product of the chain, and getting its share in the profit.

Finally, the Belt and Road transportation routes can be profitable only if relevant global value chains (as big meta-companies) will be competitive in the world market.

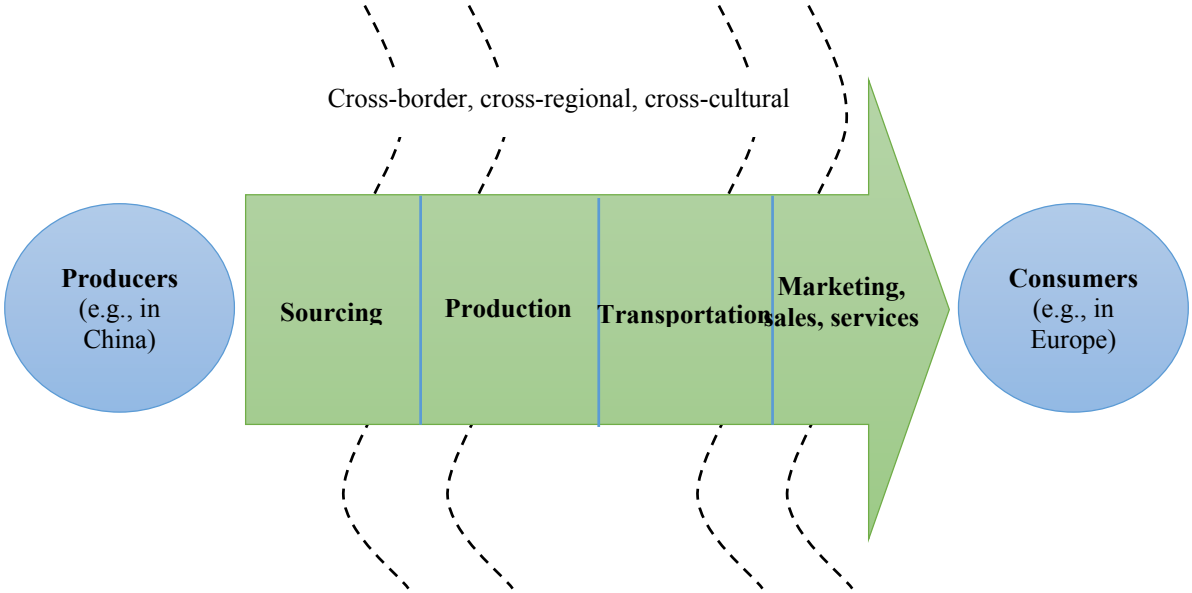


Fig. 4 Global value chain model along the Belt and Road

V. Digital Marketing Infrastructure

In the modern world all marketing activities listed above cannot be implemented without using broad opportunities given by digital marketing and ecommerce.

Digital marketing can be defined as "the application of the Internet and related digital technologies in conjunction with traditional communications to achieve marketing objectives" [10]. In other words, "digital marketing can be simply defined as achieving marketing objectives through applying digital technologies" [10].

Authors think there is a huge need of digital marketing infrastructure all along the Belt and Road (see Fig. 5).

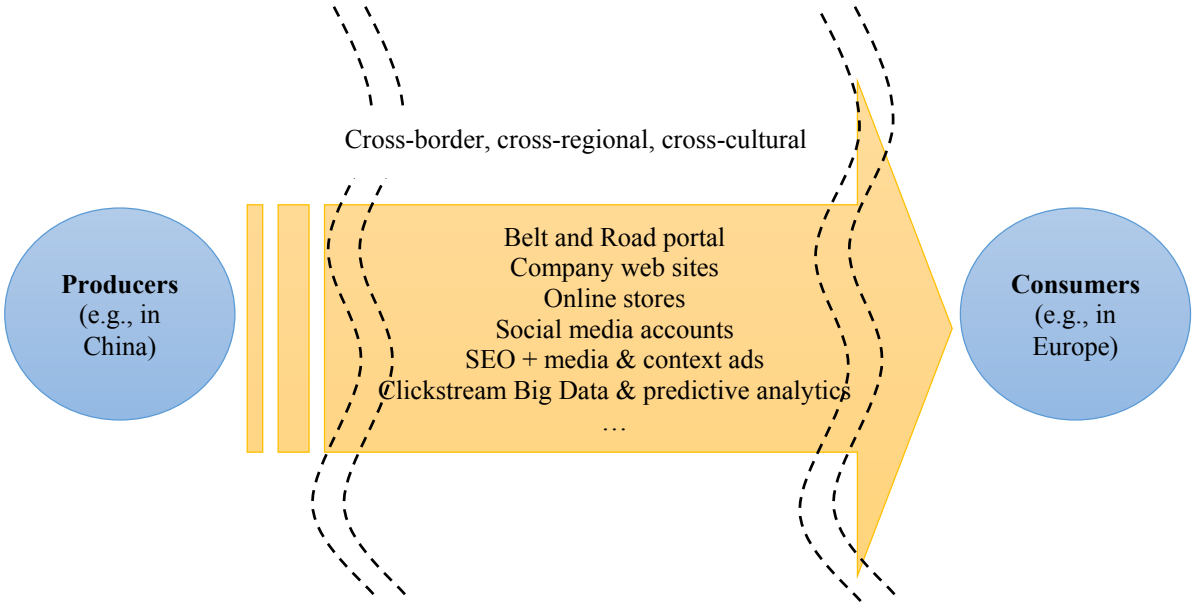


Fig. 5 Digital marketing infrastructure along the Belt and Road

To our mind, digital marketing infrastructure should include at least:

• the Belt and Road portal – not just as a source of official documents and news, but as an interactive platform where a company could find partners in order to get integrated into the most suitable value chain.

• Companies-participants' web sites – with product catalogues, contact information, online order forms.

• Online stores – predominantly oriented to the end consumers.

• Social media accounts and social media marketing (SMM) activities – for the Belt and Road in general and for a specific company, brand or product.

• Search engine optimization (SEO) activities – in order to make fully visible all elements mentioned above in a web search engine's unpaid results.

• Media & context advertisement, email marketing – as promotional tools for web sites, online stores, social media accounts, featured events, etc.

• Integrated CRM and e-CRM systems – as well as consumer clickstream Big Data & predictive analytics (e.g., [10], [11]).

The most crucial point here is the full integration of different elements of the digital marketing infrastructure along the certain global value chain along the Belt and Road.

TABLE I
KEY ELEMENTS OF MARKETING STRATEGY AT DIFFERENT LEVELS OF THE BELT AND ROAD INITIATIVE

Level of strategy	Goal	Target audience	Value proposition / marketing message	Digital tools
Belt and Road Initiative	To promote the Initiative and attract participants from different countries and regions	Investors, member-countries, mass media, general public etc.	Trade between Asia and Europe is profitable for every participant	Web portal of the Initiative, online publications, social media marketing etc.
Value chain	To form a viable business model with high synergetic effects in all parts of the value chain	Transport operators, producers, wholesalers, retailers, banks etc.	Combination of different companies from different regions based on their competitive advantages and core competencies forms a viable business model	Integrated CRM and e-CRM systems along the value chain, integrated ecommerce infrastructure
Company	To fit the global (regional) value chain in the most competitive way	End consumers, partners, intermediaries etc.	Better product or services for partners or end consumers	Company web site, online-store etc.
Belt and Road Initiative	To promote the Initiative and attract participants from different countries and regions	Investors, member-countries, mass media, general public etc.	Trade between Asia and Europe is profitable for every participant	Web portal of the Initiative, online publications, social media marketing etc.

VI. INTEGRATED MARKETING STRATEGY ALONG THE BELT AND ROAD

Therefore, in order to make the Belt and Road fully functional and successful, apart from technical and logistic issues, it is necessary to take into the account the need of marketing strategy along the Belt and Road.

Authors find it useful to distinguish in this context three main levels of marketing strategy (see Table 1), namely:

• marketing for the Belt and Road Initiative as such – for attracting investors, member-countries, and forming its positive image among mass media and general public;

• strategic marketing efforts of designing competitive global (regional) value chains along the Belt and Road by combining different companies from different regions based on their competitive advantages and core competencies in order to form viable business model;

• specific marketing for a member of global value chains along the Belt and Road in

accordance to the chosen general marketing strategy, push or pull, and the need of managing the uninterrupted all types of flows in the channel.

Integration of such multi-level marketing efforts and activities would make the Belt and Road Initiative fully functional, robust and sustainable.

VII. Special case: marketing strategy for the China-Belarus industrial park "Great Stone"

A special case that authors want to highlight is the China-Belarus industrial park "Great Stone".

This park is a territorial entity with a special legal status beneficial for doing business. The industrial park is located (see Fig. 6) in Smolevichy district, Minsk region, 25 km from Minsk, the capital of Belarus, and occupies an area of 91,5 sq. km. Geographical location is characterized by close proximity to the Minsk International Airport, the railway, the Berlin-Moscow transnational highway, as well as by access to the port of Klaipeda on the Baltic Sea which is of 500 km away [13].

Any company regardless of country of capital origin can act as a resident of the industrial park. Officially stated benefits for such a resident include the following [14], [15]:

- Products produced in the park are within the Customs Union of Belarus, Russia, Kazakhstan, Armenia and Kyrgyzstan, which has more than 183 mln people. In this region, customs duties and economic limitations are not applied;
- Access to the Eurasian Economic Union means free traffic of goods, services, capital, labor force, harmonization of legislation, technical and regulatory standards and rules;
- Commonwealth of Independent Countries (CIS) Free Trade Agreement: free trade with 11 CIS countries;
- The proximity to the neighboring European countries means reduction of transport and time expenditures when exporting goods to Europe;
- Huge tax and customs preferences (including 0% rates for most taxes until 2062).

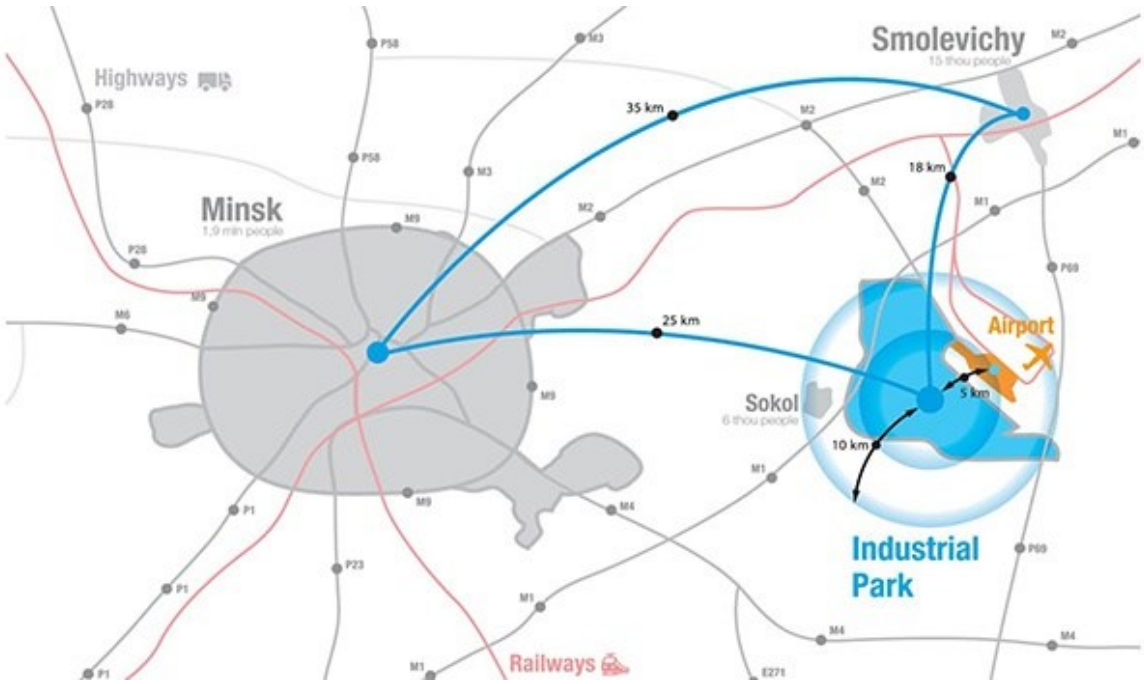


Fig. 6 The geographic location of the "Great Stone" near Minsk city [16]

Listed benefits are based on three main factors: geographical location of the Park, special legal regime given by the government, and Belarus status as a member of the Customs Union, the Eurasian Economic Union and CIS Free Trade Agreement. All this makes the Park to be a "gate" to EU and EEU markets (see Fig. 7) from the perspective of Chinese producers.

Fig. 7 The "Great Stone" position as a "gate" to EU and EEU markets

Marketing strategy for the Park should be in line with this principal vision, while concentrating on the promotion of the Park itself and its residents.

VIII. Conclusions

1) From the marketing point of view the Belt and Road should be understood not only in technical way as a transportation infrastructure, but as marketing channels from China to the rest of Eurasia and Africa and vice versa.
2) The main emphasis should be moved from the transportation goal (from the A point to the B point) to the final goal of any marketing channel – a purchase, and through a purchase to long-term win-win relationship and partnership. This requires managing, in addition to physical transportation of a product, all types of flows in the channel, including the promotional flow, which could be done either through push or pull classical strategies.
3) The concept of global value chains implies the Belt and Road to be understood as a set of multiple global value chains across Eurasia. There is a strategic marketing problem for any participant of such global value chains along the Belt and Road to find its "place" within a specific chain thus using its competitive advantages, adding value to the final product of the chain, and making profit.
4) In the modern world, marketing cannot be implemented without using broad opportunities given by digital marketing and e-commerce. Authors think there is a huge need of digital marketing infrastructure all along the Belt and Road, including the interactive platform, where a company could find suitable partners, online stores, relevant SEO and SMM activities, etc. The most crucial point here is the full integration of different elements of the digital marketing infrastructure along the certain global value chain along the Belt and Road.

5) In order to make the Belt and Road Initiative fully functional, robust and sustainable there is a need of integration of marketing efforts and activities within several levels of the project.

ACKNOWLEDGMENT

This research project was conducted for and got one of the first category prizes at The Belt and Road Initiative International Special Event Contest of Challenge Cup 2017 held at Shanghai University.

Authors want to thank Dr. Helena Zhivitskaya, Vice-Rector of BSUIR, for the initiation and supporting this research project, Dr. Larisa Arkhipova, Associate Professor of Economics department at BSUIR, for professional advice and insightful suggestions, and Organizing Committee of the 15th Challenge Cup National College Students' Extracurricular Academic Science and Technology Works Contest for the opportunity to participate and kind care at all stages of the competition.

Special thanks to Prof. Dr. Aliaksei Bykau from Belarusian State Economic University for his role in introduction the concept and importance of global value chains to the authors of this article. His (in coauthorship) book *"Trade in value added: sources of balanced economic growth"* (2017, in Russian) was one of the most influential sources for theoretical framework construction for this project.

Authors also thank two anonymous reviewers for their valuable comments, which helped to give more clarity to the text and to outline directions for future research.

REFERENCES

[1] "Belt and Road Portal", https://eng.yidaiyilu.gov.cn/

[2] "The Belt and Road Initiative - A road map to THE FUTURE", http://beltandroad.hktdc.com/en/belt-and-road-basics

[3] "Vision and Actions on Jointly Building Silk Road Economic Belt and 21st-Century Maritime Silk Road", http://en.ndrc.gov.cn/newsrelease/201503/t20150330_669367.html

[4] "Backgrounder: Economic corridors under Belt and Road Initiative", http://www.globaltimes.cn/content/1046027.shtml

[5] "The Belt and Road Initiative", http://china-trade-research.hktdc.com/business-news/article/The-Belt-and-Road-Initiative/The-Belt-and-Road-Initiative/obor/en/1/1X000000/1X0A36B7.htm

[6] Rosenbloom B., "Marketing Channels: A Management View", Fort Worth [u.a.]: Dryden Press, 1999.

[7] Kotler P., Keller K. L., "Marketing management", Upper Saddle River. N.J: Pearson Prentice Hall, 2009.

[8] Wagner G. Channel Management (Lecture, Summer Term, 2011). http://www.wiwi.uni-siegen.de/marketing/lehre/vorlesungen/channelmanagement/ss2011/channel_management_sose2011_t3.pdf

[9] "The Global Value Chains Initiative", https://globalvaluechains.org/concept-tools

[10] Chaffey D., Ellis-Chadwick F., "Digital marketing", Pearson Higher Ed, 2012.

[11] Parkhimenka U., Tatur M., Zhvakina A., Heuristic approach to online purchase prediction based on internet store visitors classification using data mining methods. In Information and Digital Technologies (IDT), 2017 International Conference on (pp. 304-307). IEEE.

[12] Parkhimenka U., Tatur M., Khandogina O, Unsupervised ranking of clients: machine learning approach to define a "good customer", CERes Journal, Volume 3, Issue 2, 2017, pp. 10-15.

[13] "Great Stone. Industrial park. About. General information", http://www.industrialpark.by/en/general-information

[14] "Great Stone. Industrial park. Benefits. EAES", http://www.industrialpark.by/en/evrazijskoe-eknomicheskoe-soobshestvo

[15] "Great Stone. Industrial park. Benefits. Tax and customs preferences", http://www.industrialpark.by/en/economic/3-benefity

[16] "Great Stone. Industrial park. Benefits. Geography", http://www.industrialpark.by/en/geography

Software Complex of Automation Examination of Seed

Anna Zhvakina, Mikhail Tatur, Svetlana Ostrouhova

Abstract— The article considers the features of the development of software complex of automation examination of seed in the State inspectorates of seed growing, quarantine and protection of plants.

Keywords— Examination of Seed, Software, Automation.

I. Introduction

No field can be sown, no packet of seeds can be realized on the market without the approbation of agricultural plants, without producing certificate of the quality of plant seeds and document on seeds plant. This defines the special importance of measures to determine quality of varietal and sowing properties of seeds of agricultural plants. State inspectorates of seed growing, quarantine and protection of plants provides quality control of the seeds of all crop grown in the territory of the Republic of Belarus, supervision over observance of requirements of technical regulations at all stages of the production and use of seeds. These actions perform in accordance with the Law of the Republic of Belarus 20-3 on May 2, 2013 about seed growing [3].

II. Organization of work in divisions of State inspection of seed growing, quarantine and protection of plants

Staff amount of the State Inspectorate is limited. Typically, in each district and regional inspectorates 3-4 people define varietal and sowing qualities of seeds. Most of the work is related to the conduct of expertise. Each examination is carried out according to a strictly regulated procedure [1]. The employee must always operate with a large number of reference data.

Table 1 presents some examples of requirements for varietal and sowing qualities of seeds of cereals. You can see that these requirements are divided not only for different names of agricultural plants, but also depend on the classification: varietal and sowing qualities of seeds, for parental components and for commercial purposes. There is also a separation of requirements for each stage of seed reproduction: original seeds, elite seeds, first reproduction, second-third reproductions and subsequent reproductions.

For each plant species, various impurities and different pests are considered. If we are talking about potatoes, the set of qualities is divided not only for elite and reproduction seeds, but the original seeds have different requirements: micro-plants in culture in vitro, the first tuber generation, the nursery of pre-breeding, super elite. The following qualities are evaluated:

1) Cleanliness of sort of landings,
2) Presence of plants affected by disease (by external signs), including:
   - light viral diseases (ordinary mosaic, mosaic twisting of leaves);
   - severe viral diseases (wrinkled mosaic, banded mosaic, twisting of leaves);

A. Zhvakina, Intellectual Processors Ltd., Minsk, Belarus (e-mail: antim07@mail.ru)
M. Tatur, Intellectual Processors Ltd., Minsk, Belarus (e-mail: tatur@i-proc.com)
S. Ostrouhova, Electronic Computing Machines Department, BSUIR, Minsk, Belarus (e-mail: ostrouhova@bsuir.by)

- soil viruses;
- viroids;
- bacterial (black leg).

The presence of plants affected by a viral and bacterial infection in a latent form, including:

- viruses X, S, M, Y, L, A;
- bacterial infection (black leg).

3) Size of tubers (for varieties with an elongated shape, for varieties with a round-oval shape), mm.
4) Presence of tubers that do not fit in size.
5) Appearance of the plant.
6) Presence of plants, overgrown, other botanical varieties, pcs.
7) Presence of tubers with signs of "suffocation", frozen, with burns, ugly, cut, crushed, with peeled skin more than 1/3 of the tuber surface.
8) Presence of tubers with mechanical damages.

TABLE I
REQUIREMENTS FOR VARIETAL AND SOWING QUALITIES OF SEEDS OF CORNY AND CEREALS OF AGRICULTURAL PLANTS

Name of the agricultural plants	Quantity	Stages of Seed Reproduction
Oats (film forms)	Requirements for varietal qualities of seeds: High purity Infestation in sowing Requirements for seed quality of seeds: Purity of seed The content of seeds of other cultivated plants, except for winter rye seeds, pcs / kg The content of seeds of weed plants, total pcs / kg, including: difficult-to-separate (oats), poisonous (heliotrope pubescent, trihodessma gray-haired) Admixture of pouch bags and their parts Admixture of ergot sclerotia Germination Humidity Population of living pests and their larvae, except for mites pcs / kg Availability of mite pcs / kg	original seeds elite seeds first reproduction second-third reproductions subsequent reproductions
Varieties of Maize hybrids	Requirements for varietal qualities of seeds for parental components and for commercial purposes: Sort typical: Seed Hybridity Level: simple hybrids, double interlinear and three linear hybrids Contents of xenia grains: pcs. Requirements for quality of seeds Purity of seeds, not less than Seed content of other cultivated plants (does not apply to seeds intended for sowing, for fodder purposes), pcs / kg The content of weed seeds, pcs / kg including poisonous	original seeds elite seeds first reproduction second-third reproductions subsequent reproductions
Amaranth (Annual fodder and honey grasses)	Requirements for seed quality of seeds Purity of seeds, Seed content of other types of annual fodder and herbs (not for seeds intended for fodder sowing purposes), The content of weed seeds, including the most harmful (Cirsium setosum M. B.), pcs / kg Occupancy of live pests and their larvae, pcs / kg Germination, Humidity,	original and elite seeds, reproduction seeds
Origanum vulgare L. (seeds of vegetable agricultural plants)	Requirements for varietal qualities of seeds (seed,): High purity, The admixture of varieties and sharp hybrids among the total admixture in the third category, Requirements for seed quality of seeds: Purity of seeds, Seed content of other plants, including weed seeds Population of mite, live pests and their larvae, pcs / kg Germination, Humidity	original seeds elite seeds first reproduction second-third reproductions subsequent reproductions hybrid

9) Presence of tubers with pest damage.
10) The presence of tubers affected by disease, total, including:
- wet rot;
- black foot;
- dry rot (fusarium, anthracnose);
- stem nematode;
- ordinary and silvery scab (damage more than 1/3 of the tubers surface);
- Rhizoctonia (with damage from 1/10 to 1/4 inclusive the surface of the tubers).

11) Presence of land and impurities.
12) The presence of plants, tubers affected by a viral and bacterial infection in a latent form.

Thus, it can be seen what a huge amount of diverse data needs to be taken into account in order to conclude that a certain material is suitable.

Every action is accompanied by a large number of current registration and reporting documents [2].

Currently, there are not automation of all kinds of work in all divisions of State inspection of seed growing, quarantine and protection of plants. Results are recorded manually, stored on paper. As a result, the same data have to be put in different types of magazines. To generate reports, you must use a variety of source documents. It increases the time of the decision and reduces the efficiency of the inspection as a whole. As a consequence, a significant part of working time of experts have to spend on unproductive work, making of documents, reports, formularies, etc. Especially this situation is exacerbated in times of increasing congestion in State inspection, for example, in the period before planting.

Thus, it is necessary to increase the efficiency of the labor of specialists departments of test the State inspection of the Republic of Belarus on seed-growing, quarantine and protection of plants (of 6 regional and 129 district inspections), through the development and implementation of software system for automating the examination of seed.

III. The business process scheme formalized

Figure 1 shows a formalized model of seeds expertise in local and regional inspections realized in the developed software package.

The **Applicant** shall submit to the **Inspection** a statement which indicates information about themselves, about analyzable agricultural plant and the necessary test types (name of work), receives an invoice, formed on the basis of statements, pays bills and provides a document confirming the payment to the inspection, provides the material for analysis, will receive a document of the results after analysis.

The **Inspection** has ability to make changes to the base price list (limited access) according to the data received from bookkeeping department. **Inspection** makes out Order based on the statement of the **Applicant**, forms the invoice, receives the document on payment and material for the analysis, directs material to the analysis and issues documents on results of analysis. In case of lack of payment of the Order, **Inspection** deletes the Order from the list of orders. **Inspection** has access to the list of the available and executed orders.

Analyst selects samples, conducts tests, records data in electronic forms and forms all types of reporting documentation.

After the automation of the examination of seed in the State inspection on seed-growing, quarantine and protection of plants, the following approach to order-taking procedure implemented:

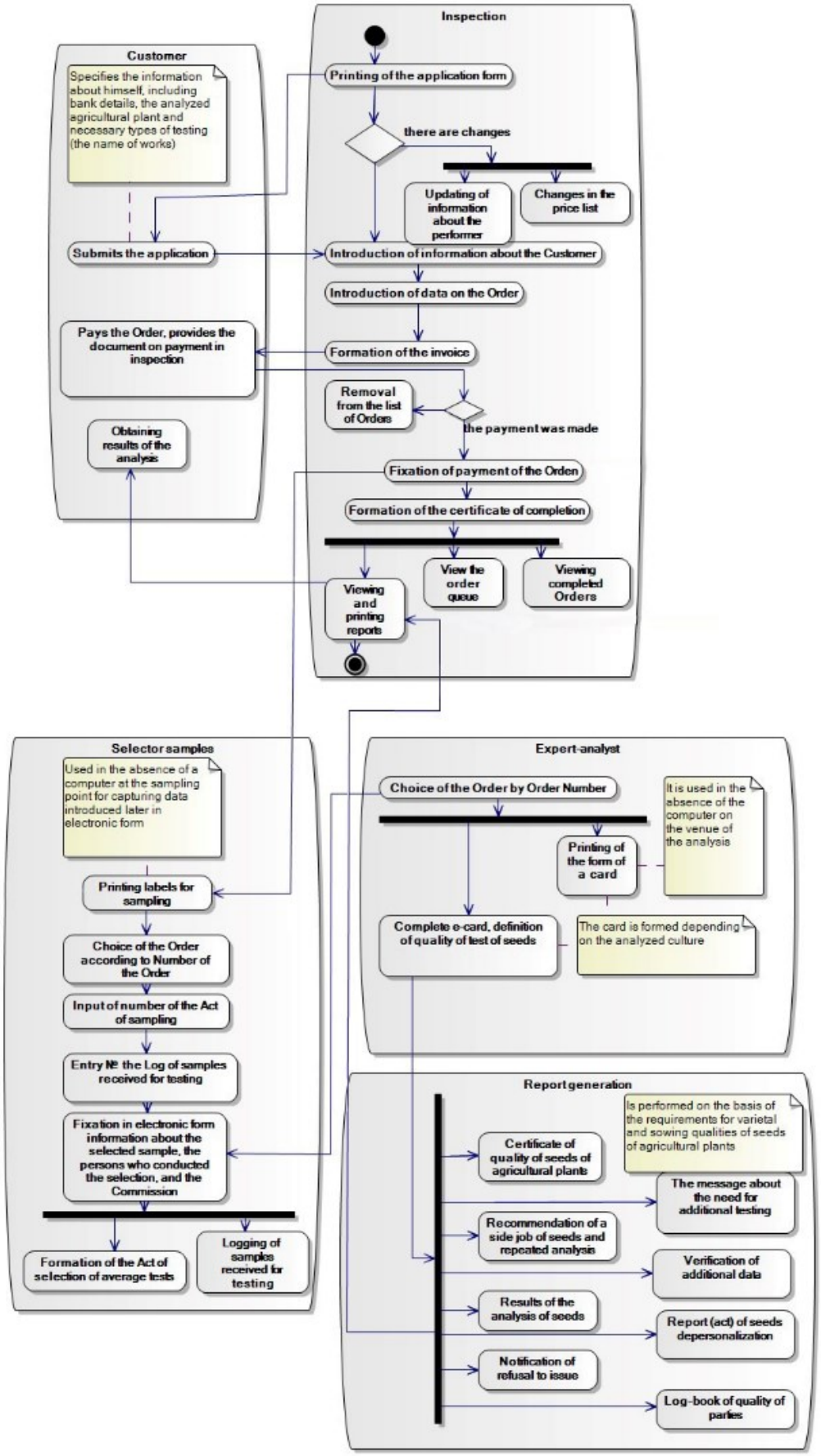


Fig. 1 Formalized model of seeds expertise in local and regional inspections

- Updating of information about the performer;
- Introduction of information about the Customer;
- Introduction of data on the Order;
- Changes in the price list;
- Formation of the invoice;
- Formation of the act of the performed works.

“Sampling for tests” is made of the material directed to the analysis, which provides information about the samples. Based on this information and the information received at the reception of the order Act of average sampling and sample Magazine, received at the trial are automatically generated. Upon transition, the dialog box with a reminder on need to save data will be brought out of the Order and Sampling tabs.

Use “Sample Analysis” option includes:

- The “Filling of a card” which is carried out in process of carrying out the analysis;
- “Formation of the report”, which can be represented as the following: “Notice of refusal to issue”, “Recommendation on seed refinement and re-analysis”, “Certificate of Quality of agricultural plants seeds”, “The results of the analysis of seeds”;
- “The formation of conclusion”, expandable up to “The message on need of additional tests” and “Verification of additional data”.

By results of tests, the following options of use are allocated:

- “Formation of the report on a depersonalization of seeds” which is submitted by the Act of a depersonalization;
- "Formation of the Journal of Accounting for Party Quality", which can be extended to "Forming Reports of Arbitrary Form".

Figure 2 shows an example of the program interface for an order for sampling from lots of seeds of agricultural plants and for their analysis.

The system is capable of supporting a minimum of 15 concurrent users associated with a common database.

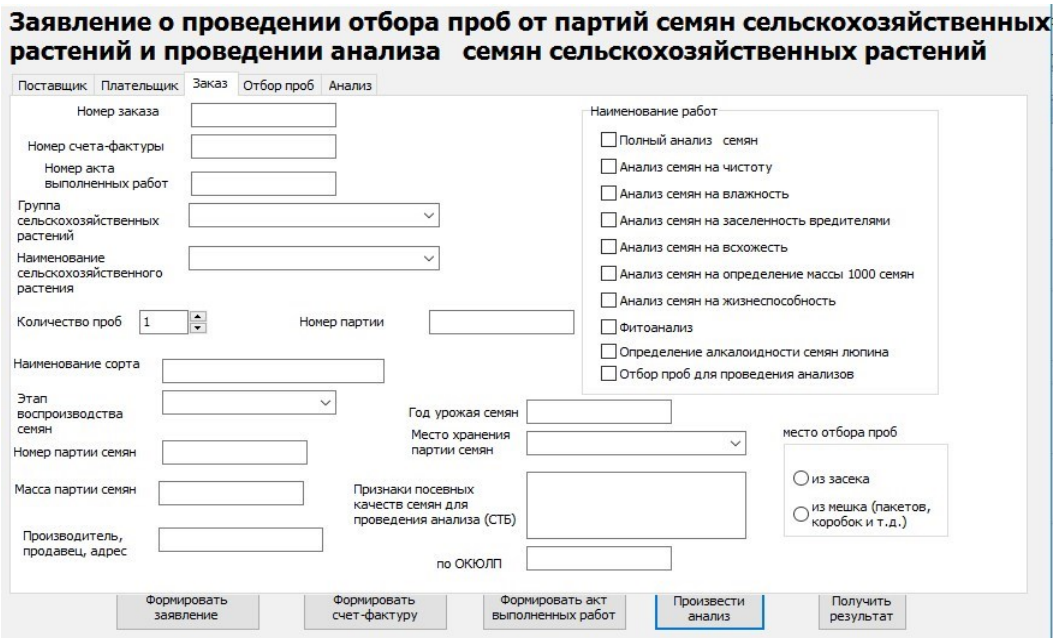


Fig. 2 The program interface of Automation Examination of Seed

The response time for typical tasks is no more than 5 seconds, for complex tasks no more than 20 seconds.

AWP is available on working days during working hours (usually from 8 to 18, unless otherwise specified by the order for the enterprise).

The time spent on maintenance of the system does not exceed 3% of the total operating time.

At the core of the system is an industrial database of relational access.

All access to information is carried out through the ODBC driver. C # programming language was used for development.

The developed software has the following advantages:

- Initial specialization in the subject area (as opposed to a widespread universal system "1C" type) increases usability;
- Modules for integration with existing systems, cross-sectorial document flow of Republic of Belarus;
- Flexibility and survivability (will allow increasing functionality of system);
- Openness (maintainability);
- Service automation (remote setting and maintenance of the system, back-up and optimization of stored data);
- Possibility of simultaneous use by multiple users;
- Usability (the possibility of rapid development of the user interface specialists with low levels of computer literacy, availability of contextual background information, enclosure of menu items no more than three levels, scaling for people with the weakened sight);
- Documents developed by system will meet the requirements of national standards, regulations and orders;
- Access to system will be provided by means of several preset automated workplaces (AWP), each of which has a certain set of the rights for viewing, creation, editing or other actions with documents;
- Scalability in terms of adding new workstations. Last two points are intended for simplification by reducing the number of options available at the same time and improve the reliability of differentiation by user authorization.

IV. Conclusions and future work

Using the software system of automation of examination of seeds in one State inspectorate for seed farming, a quarantine and protection of plants improved the efficiency and quality of work of this organization at the expense of an increase in labor productivity on 15 %.

As a result, time of examination is reduced, the quantity of tests increase; there is ability to quickly develop the necessary reports on any kind of activity of individual inspections, and in general for the Republic of Belarus. In a future more time will be allocated directly to conducting examinations, which will improve their quality. This promotes the formation of seed only from carefully selected specimens, which in turn increase the productivity of plants and, consequently, the economic security of our country.

In social terms, the effect of the use of this software will appear to facilitate routine work, often low paid. A possible reduction in the cost of these services for agricultural producers and importers will increase accessibility for the public inspection services.

The information system can be used in any State inspectorate for seed farming, a quarantine and protection of plants in the territory of Republic of Belarus as it is autonomous, and with use of the centralized access to the Republican database for storage of results of work of the State inspections.

References

[1] Regulation of the Ministry of Agriculture and Food of the Republic of Belarus of October 29, 2015 № 37 "On establishing the requirements for varietal and sowing characteristics of seeds of agricultural plants".

[2] Decree of the Ministry of Agriculture and Food of the Republic of Belarus 19 November 2013 number 61 "On the establishment of forms of documents".

[3] The Law of the Republic of Belarus of May 2, 2013 № 20-W "On the seed growing".

CERES support the research project. Centers of Excellence for young RESearchers (Reg.no. 544137-TEMPUS1-2013-SK-JPHES),

Co-funded by the Tempus Programme of the European Union

A Fuzzy Information Density Based Clustering Algorithm in Medical Data Analysis

V. Ponomarenko

***Abstract* —** The paper is devoted to the transformation of numerical values into linguistic values in medical data analysis. A software tool, based on clustering, was developed for this purpose. The implementation results are presented in this paper. A modification of the fuzzy clustering algorithm FEBFC is also introduced, which supposes using a fuzzy information density instead of fuzzy entropy. The proposed algorithm is compared with several well-known clustering algorithms on medical data sets.

Keywords — cluster validity indices, fuzzification, clustering algorithms, fuzzy clustering, medical data analysis.

I. Introduction

During the last few decades, the face of the modern world was qualitatively changed by information technologies. Various technical means and information delivery channels, based on progressive information and communication systems, have revolutionized human life, that nowadays is inseparably associated with huge flows of data. These data are involved in all spheres of human activity – social, economic, political, spiritual. However, the information contained in such data can be whether very valuable or completely useless. Extracting various kinds of useful information from the data sets is one of the main tasks of the modern science called data mining.

Data mining proposes methods, that today are widely used in healthcare, which is in one of the fundamental fields of the social sphere of modern life. Such popularity of the data mining methods was formed mainly due to the rapid development of medical devices and therapy technologies, that allow to produce and to store large amount of data. Producing of new data is mostly achieved in a process of providing medical services. For example, imagine a patient who came to a polyclinic to examine his digestive system. A medical worker, using a special probe or ultrasound device, measures necessary medical indicators and records them in a medical report. Obtained in a such way medical data are usually persisted in some database for possible using in future. Therefore, storing medical data is achieved through the use of various database systems.

The information contained in medical data is extremely important for solving diagnostic, therapeutic, statistical, administrative and other tasks in the field of medicine, e.g. determination of a correct treatment, definition of a patient's group of risk and prevention of diseases. Solving of these tasks has a huge impact on a quality of medical services, life expectancy, mortality and time of illnesses of population.

In the field of medicine, both numeric (continuous) and nominal (linguistic) types of data are used. The numeric data type is used for representing value of a continuous medical indicator, e.g. age of a patient, body mass index, resting blood pressure, albumin and globulin ratio. Variables of the nominal data type usually keep a name of some state of a categorical medical indicator, e.g. a patient's appetite can be good or poor, a tumor can be malignant or benign. In medical data analysis obtaining a continuous value is often not enough informative to make a

V. Ponomarenko, University of Žilina, Žilina, Slovakia (e-mail: volod.ponomar@gmail.com).

conclusion about patient's state for determining necessary treatment, so the medical worker has to associate this value with an appropriate nominal value. This is a typical situation when a transformation from numeric into linguistic values comes into play. The numeric values of an attribute obtained as a result of the transformation can be relatively easy converted into fuzzy data specified by a membership function. The whole process of obtaining fuzzy values from numeric values is called fuzzification.

Fuzzy medical data, obtained as a result of fuzzification, are very valuable. They can be used, in particular, for increasing the healthcare system reliability through the reducing potential medical failures, that is discussed in papers [1]–[2].

The process of transformation from numeric into linguistic values, that is based on cluster analysis, is the subject of study in this paper. Various of clustering algorithms, that perform cluster analysis, can be found in literature. This variety includes fuzzy clustering algorithms, that implies belonging of an object to several clusters simultaneously. A promising clustering algorithm Fuzzy Entropy Based Fuzzy Classifier (FEBFC), proposed in paper [3], assumes using a fuzzy entropy, based on Shannon's entropy, as a criterion of optimality. In this paper a modification of the FEBFC algorithm is introduced, based on different approach to optimality criterion: a fuzzy information density is used instead of the fuzzy entropy measure. This study also includes development of a software for transformation values of any numeric attribute of the medical data set into fuzzy values, based on clustering algorithms. Several fuzzy clustering algorithms are implemented in this software solution as well as the mentioned FEBFC algorithm and its modification. Implementation details and accuracy comparison of clustering algorithms are discussed below in this paper.

II. Fuzzy Information Density Based Fuzzy Classifier

The mentioned above FEBFC clustering algorithm was proposed by Hahn-Ming Lee, Chih-Ming Chen et al. [3]. According to the proposed approach, the fuzzy entropy $FE(\tilde{A})$ is defined on the universal set $X = \{r_1, r_2, \dots, r_n\}$, where $i = 1, 2, \dots, n$, for the elements within an interval (cluster) in a non-probabilistic way:

$$FE(\tilde{A}) = \sum_{j=1}^{m} FE_{C_j}(\tilde{A}) = \sum_{j=1}^{m} -D_j \log_2 D_j \tag{1}$$

where $\tilde{A}$ is a fuzzy set defined on an interval of pattern space which contains k elements ($k < n$); $C_1,\ C_2, \dots, C_m$ represent m classes into which the n elements are divided; $FE_{C_j}(\tilde{A})$ is the fuzzy entropy of the elements of class j in an interval, defined as $FE_{C_j}(\tilde{A}) = -D_j \log_2 D_j$; D_j is the match degree with fuzzy set $\tilde{A}$ for the elements of class j in an interval, where $j = 1, 2, \dots, m$, defined as $D_j = \frac{\sum_{r \in S_{C_j}(r_n)} \mu_{\tilde{A}}(r)}{\sum_{r \in X} \mu_{\tilde{A}}(r)}$; $\mu_{\tilde{A}}(r_i)$ is the mapped membership degree of the element r_i with the fuzzy set $\tilde{A}$; $S_{C_j}(r_n)$ is a set of elements of class j on the universal set X (subset of the universal set X).

The FEBFC algorithm assumes using a fuzzy entropy cluster validity index (I_{FE}), also known as a total fuzzy entropy, for determining an optimal number of clusters. The index is defined as a sum of fuzzy entropies of all clusters:

$$I_{FE} = \sum_{i=1}^{k} FE_i^* \tag{2}$$

where k – the number of clusters; FE_i^* – the fuzzy entropy of the i-th cluster, calculated as a sum of fuzzy entropies of all fuzzy sets on the i-th interval.

The proposed cluster validity index would be perfect if all intervals had equal length and quantity of patterns on them. But usually there are several clusters of different size among one data set. Distances between the patterns are also different. Thus, a simple addition of fuzzy entropy values of intervals may lead to inaccurate results. Alternatively, the fuzzy information density measure can be used instead of the fuzzy entropy measure as an optimality criterion, that will potentially lead to more accurate clustering results, because the fuzzy information density takes cluster sizes into account.

The information density measure was defined in paper [4]. Based on it, of the fuzzy information density FD_q of the q-th cluster ($q = 1,2,\dots,k$) is defined as:

$$FD_q = \begin{cases} FE_q/\log_2(n_q + 1), & \text{if } n_q > 0 \\ 0, & \text{if } n_q = 0 \end{cases} \tag{3}$$

where FE_q – the fuzzy entropy on the q-th interval; n_q – the number of patterns on the q-th interval.

The fuzzy entropy cluster validity index (I_{FE}) should be then replaced by the fuzzy information density cluster validity index (I_{FD}), that is also called the total information density. It is defined as:

$$I_{FD} = \sum_{q=1}^{k} \omega_q \times FD_q \tag{4}$$

where FD_q – the fuzzy information density on the q-th interval; $\omega_q = \frac{n_q}{n}$ is a weight coefficient; n – the number of patterns in the data set; n_q – the number of patterns on the q-th interval. Then the optimal number of clusters k^* is calculated as

$$I_{FD}(k^*) = \min_{2 \le k \le n-1} I_{FD}(k) \tag{5}$$

As a result of applying the proposed changes to the original FEBFC clustering algorithm, a new algorithm was obtained. It was called a Fuzzy Information Density Based Fuzzy Classifier (FIDBFC).

The FIDBFC clustering algorithm consists of the following steps:

Step 1. Set the initial number of clusters (intervals) $k \coloneqq 2$.

Step 2. Locate the centers of intervals using following subsequence of steps:

2A. Find the initial centers of intervals $c_1, c_2, \dots, c_k$ using formula:

$$c_q = x_{min} + (x_{max} - x_{min}) \times \frac{q-1}{k-1}, \qquad q = 1,2,\dots,k.$$

2B. Assign each element of the distribution to a corresponding interval with the smallest Euclidian distance to the interval center:

$$|x_i - c_q^*| = \min_{1 \le q \le k} |x_i - c_q|$$

where c_q^* is the closest center to the element x_i.

2C. Recompute the cluster centers.

$$c_q = \frac{\sum_{i=1}^{n_q} x_i^q}{n_q}$$

where n_q is the total number of patterns x_i^q, that belong to q-th cluster.

2D. Compare recomputed cluster centers with previous. If any center was changed then go to *Step 2B*. Otherwise, go to *Step 3*.

Step 3. Assign the membership function for each interval according to:

$$\mu_1 = \begin{cases} 1, & \text{for } x \le c_1 \\ \frac{c_2 - x}{c_2 - c_1}, & \text{for } c_1 < x \le c_2 \\ 0, & \text{otherwise} \end{cases} \qquad \mu_q = \begin{cases} 0, & \text{for } x \le c_{q-1} \\ \frac{x - c_{q-1}}{c_q - c_{q-1}}, & \text{for } c_{q-1} < x \le c_q \\ \frac{c_{q+1} - x}{c_{q+1} - c_q}, & \text{for } c_q < x \le c_{q+1} \\ 0, & \text{otherwise} \end{cases}$$

$$\mu_k = \begin{cases} 0, & \text{for } x < c_{k-1} \\ \frac{x - c_{k-1}}{c_k - c_{k-1}}, & \text{for } c_{k-1} \le x < c_k \\ 1, & \text{otherwise} \end{cases}$$

where $q = 2,3,\ldots,k-1$.

Step 4. Compute the $I_{FD}(k)$ for k clusters and $I_{FD}(k-1)$ for $k-1$ clusters according to formula (5).

Step 5. If $I_{FD}(k) < I_{FD}(k-1)$, then partition again ($k := k+1$) and go to *Step 2*; otherwise, $k-1$ is the optimal number of clusters.

For illustrating the FIDBFC algorithm, we use the following example. Let X be a distribution of three classes of objects represented by values of some attribute of these objects. The distribution divided into three and four intervals is shown in Figure 1 (a) and (b) respectively as a set of objects △, □ and ○, placed on x axis. Position on the axis corresponds with a value of the attribute.

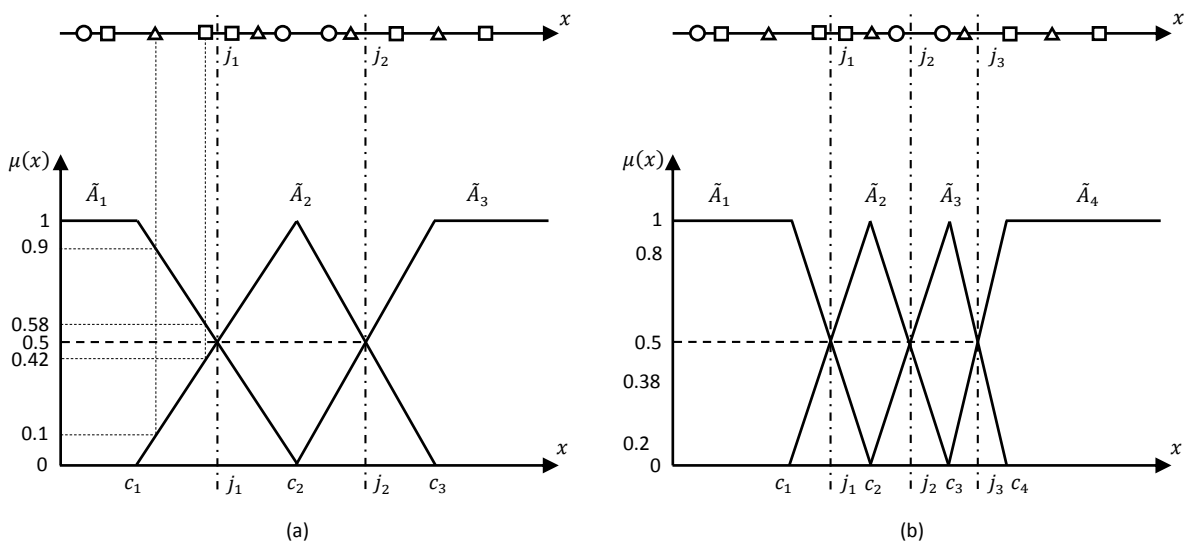


Fig. 1. Example of a distribution of 3 classes of objects (△, □ and ○ denote class 1, class 2 and class 3 respectively) with corresponding membership functions

In the first considered case (Figure 1 (a)), the distribution is divided into three intervals, that

are $(-\infty; j_1)$, $[j_1; j_2)$ and $[j_2; \infty)$. On these intervals fuzzy sets $\tilde{A}_1$, $\tilde{A}_2$ and $\tilde{A}_3$ are obtained using a membership function. The centers of fuzzy sets are denoted as c_1, c_2, c_3. Let us introduce a calculation process of the total fuzzy entropy of the distribution with mentioned dividing.

In the below paragraphs the sequence of steps needed to calculate the fuzzy entropy measure of the interval $(-\infty; j_1)$ is detailly described. The measure calculation process for other intervals is skipped, because the same method is used.

At first, we find a total membership degree for each class on values of the fuzzy set $\tilde{A}_1$ from the interval:

- total membership degree of "$\triangle$" is 0.9;
- total membership degree of "$\square$" is $1 + 0.58 = 1.58$;
- total membership degree of "$\circ$" is 1.

Then we calculate match degrees:

- $D_\triangle = 0.9/(0.9 + 1.58 + 1) = 0.9/3.48 = 0.25862$;
- $D_\square = 1.58/3.48 = 0.45402$;
- $D_\circ = 1/3.48 = 0.28736$.

In the next step we calculate fuzzy entropies of $\tilde{A}_1$ on $(-\infty; j_1)$:

- $FE_\triangle(\tilde{A}_1) = -0.25862 \times \log_2 0.25862 = 0.50459$;
- $FE_\square(\tilde{A}_1) = -0.45402 \times \log_2 0.45402 = 0.51721$;
- $FE_\circ(\tilde{A}_1) = -0.28736 \times \log_2 0.28736 = 0.51698$;
- $FE_1(\tilde{A}_1) = FE_\triangle(\tilde{A}_1) + FE_\square(\tilde{A}_1) + FE_\circ(\tilde{A}_1) = 1.53878$.

Similarly, the fuzzy entropies of $\tilde{A}_2$ and $\tilde{A}_3$ on $(-\infty; j_1)$ are calculated:

- $FE_1(\tilde{A}_2) = FE_\triangle(\tilde{A}_2) + FE_\square(\tilde{A}_2) + FE_\circ(\tilde{A}_2) = 0.70627$;
- $FE_1(\tilde{A}_3) = FE_\triangle(\tilde{A}_3) + FE_\square(\tilde{A}_3) + FE_\circ(\tilde{A}_3) = 0$.

The fuzzy entropy of the interval $(-\infty; j_1)$ equals:

- $FE_1^* = FE_1(\tilde{A}_1) + FE_1(\tilde{A}_2) + FE_1(\tilde{A}_3) = 2.24505$.

Similarly, the fuzzy entropies FE_2^* and FE_3^* of the corresponding intervals $[j_1; j_2)$ and $[j_2; \infty)$ can be obtained:

- $FE_2^* = FE_2(\tilde{A}_1) + FE_2(\tilde{A}_2) + FE_2(\tilde{A}_3) = 3.67795$;
- $FE_3^* = FE_3(\tilde{A}_1) + FE_3(\tilde{A}_2) + FE_3(\tilde{A}_3) = 0.95096$.

After obtaining the fuzzy entropy values, the information density measures can be calculated for each interval:

- $FD_1 = FE_1^*/\log_2(k+1) = 2.24505/\log_2 5 = 0.96689$.
- $FD_2 = FE_2^*/\log_2(k+1) = 3.67795/\log_2 6 = 1.42283$;
- $FD_3 = FE_3^*/\log_2(k+1) = 0.95096/\log_2 4 = 0.47548$.

Finally, according to (4) the total fuzzy information density measure of the distribution (a) is calculated in the following way:

- $I_{FD}^{(a)} = (4/12) \times 0.96689 + (5/12) \times 1.42283 + (3/12) \times 0.47548 =$
  $= 1.03401$.

In the second considered case (Figure 1 (b)), the distribution is divided into four intervals, that are $(-\infty; j_1)$, $[j_1; j_2)$, $[j_2; j_3)$ and $[j_3; \infty)$. Using the same method, as described before, the total fuzzy information density measure for this case of dividing of the distribution is calculated as:

- $I_{FD}^{(b)} = (3/12) \times 0.78318 + (4/12) \times 0.99318 + (2/12) \times 0.90356 +$
  $+ (3/12) \times 0.47122 = 0.79526$.

According to obtained results ($I_{FD}^{(b)} < I_{FD}^{(a)}$), dividing the distribution into four intervals is preferable. Comparison of clustering results of the introduced algorithm is described below.

III. A Software Tool for Cluster Analysis and Fuzzification

A significant part of the study was devoted to development of a software for transformation of values of any numeric attribute of a medical data set into fuzzy values. This software is based on fuzzy clustering algorithms and satisfies following functional requirements:

- reading data set from a file;
- basic analysis of the initial data set;
- graphical visualization of basic analysis results;
- fuzzy clustering algorithms implemented;
- importing clustering results from external software;
- graphical visualization of clustering results;
- fuzzification of the initial data set depending on clustering results;
- writing fuzzification result to a file.

According to the above list, the most of functionality of the developed software is related to fuzzy clustering. Therefore, it was called the Fuzzy Clustering Tool. The software was implemented in C++ programming language using Qt 5.10 framework. The main window of it is shown in Figure 2.

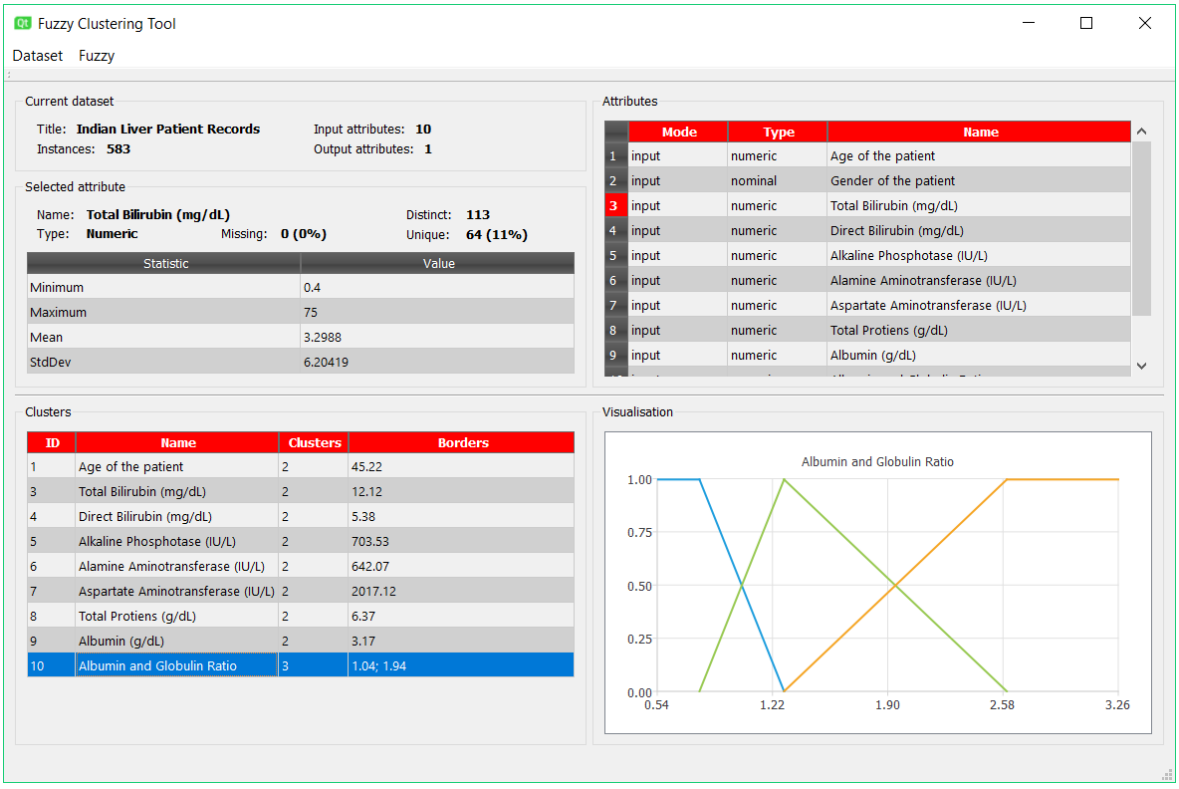


Fig. 2. The main window of the Fuzzy Clustering Tool.

IV. Comparison of the FIDBFC Algorithm with Other Fuzzy Clustering Algorithms on Medical Data

Using the implemented Fuzzy Clustering Tool, the introduced in this paper FIDBFC algorithm was compared with following fuzzy clustering algorithms: Fuzzy c-Means (FCM) [5]–[6], Gustafson-Kessel (GK) algorithm [7], Gath-Geva (GG) algorithm [8], Multi-Interval Discretization (MID) [4] and Fuzzy Entropy Based Fuzzy Classifier (FEBFC) [3]. The first three algorithms (Fuzzy c-Means, Gustafson-Kessel algorithm and Gath-Geva algorithm) need

a cluster validity index to define an optimal number of clusters, therefore the Pairing Frequency index, proposed in paper [9], was additionally implemented into the software tool.

Evaluation of fuzzy clustering results was performed through the Clustering Accuracy Indices, that can be divided into two groups: Internal indices (uses only input attributes data) and External indices (uses information about belonging of a pattern to some class of data) [10]–[11]. The Internal indices can be used in both supervised and unsupervised learning, but the External indices can be used in case of supervised learning only.

In a literature of clustering a lot of various internal indices can be found, but the most well-known of them are the following [12]–[18]:

- Partition Coefficient index;
- Partition Entropy index;
- Fukuyama-Sugeno index;
- Xie-Beni index.

Among the external indices the following can be highlighted [19]–[20]:

- Purity index;
- Normalized Mutual Information index.

The listed above clustering algorithms were evaluated and compared using these internal and external indices on the following medical data sets, obtained from the Kaggle [21] and UCI [22] Machine Learning Repositories:

- Pima Indians Diabetes;
- Heart Disease;
- Breast Cancer Wisconsin;
- Indian Liver Patient Records;
- Chronic Kidney Disease.

Table 1. Clustering Accuracy Indices calculated for the fuzzification performed on the Pima Indians Diabetes Dataset

Algorithm	Partition Coefficient index	Partition Entropy index	Fukuyama-Sugeno index	Xie-Beni index	Purity index	Normalized Mutual Information index
FCM	0.74530	0.55755	15.49595	0.13241	0.67337	0.04168
GK	0.81605	0.39656	9.10605	62.95703	0.66701	0.04180
GG	0.80680	0.42239	7.03228	129.22207	0.66321	0.03090
MID	0.96787	0.06993	9.38145	187.26315	0.65865	0.02023
FEBFC	0.84070	0.34661	5.35528	0.09274	0.65654	0.02903
FIDBFC	0.83577	0.35898	5.69637	0.09239	0.65654	0.02913

According to the obtained values of the Partition Coefficient and the Partition Entropy indices, the most accurate is the MID algorithm (see Table 1). The FEBFC algorithm is the most accurate according to the Fukuyama-Sugeno index. The FIDBFC algorithm is the most accurate according to the Xie-Beni index. The Purity indices identify the FCM as the most accurate algorithm and Normalized Mutual Information indicates the results of the GK algorithm as the best. The FIDBFC, which is the modification of the FEBFC, gives better results than the FEBFC according to the Xie-Beni index, that is one of the most relevant internal indices, and according to the Normalized Mutual Information, that is one of the most relevant external indices. Thus, modification of the FEBFC algorithm, proposed in this paper, leads to

better clustering results of the Pima Indians Diabetes Dataset according to two significant indices.

Table 2. Clustering Accuracy Indices calculated for the fuzzification performed on the Cleveland Heart Disease data set

Algorithm	Partition Coefficient index	Partition Entropy index	Fukuyama-Sugeno index	Xie-Beni index	Purity index	Normalized Mutual Information index
FCM	0.74495	0.56002	5.92440	0.11528	0.54916	0.04926
GK	0.80537	0.42032	3.17181	9.16700	0.54603	0.05179
GG	0.73040	0.58724	5.84767	62.60825	0.55168	0.04751
MID	0.98750	0.02764	4.50228	313.08871	0.54940	0.06126
FEBFC	0.81525	0.40006	2.25668	0.10192	0.54125	0.03940
FIDBFC	0.82118	0.38605	2.00213	0.10587	0.54125	0.03905

As we can see in Table 2, the MID algorithm gives the most accurate clustering results according to the Partition Coefficient, the Partition Entropy and the Normalized Mutual Information indices. Clustering results of the GG algorithm are the most accurate only according to the Purity index as well as clustering results of the FEBFC algorithm are the most accurate according to the Xie-Beni index.

The FIDBFC algorithm is the most accurate according to the Fukuyama-Sugeno index. It also better than the original FEBFC algorithm according to the Partition Coefficient and the Partition Entropy indices, has the same accuracy according to the Purity index and a bit worse according to other two indices. Thus, modification of the FEBFC algorithm, proposed in this paper, leads to better clustering results of the Heart Disease data set according to at least half of considered indices.

Table 3. Clustering Accuracy Indices calculated for the fuzzification performed on the Breast Cancer Wisconsin data set

Algorithm	Partition Coefficient index	Partition Entropy index	Fukuyama-Sugeno index	Xie-Beni index	Purity index	Normalized Mutual Information index
FCM	0,70180	0,65043	9,97471	0,17636	0,78141	0,20863
GK	0,80869	0,41571	3,11619	136,80074	0,77748	0,19883
GG	0,71908	0,61341	8,52217	30,85465	0,78417	0,20820
MID	0,91949	0,17782	8,03281	243,40205	0,76021	0,19926
FEBFC	0,83261	0,36418	3,60073	0,09625	0,75014	0,19384
FIDBFC	0,78031	0,47621	6,69028	0,10761	0,76131	0,20961

As we can see in Table 3, the MID algorithm gives the most accurate clustering results according to the Partition Coefficient and the Partition Entropy indices. According to the Fukuyama-Sugeno, Xie-Beni and Purity indices, clustering results of the GK, FEBFC and GG algorithms are the most accurate correspondingly.

The FIDBFC algorithm is the most accurate according to the Normalized Mutual Information index. It also more accurate than the original FEBFC algorithm according to the Purity index, but less accurate according to the other indices. Thus, for the Breast Cancer Wisconsin data set the FEBFC algorithm in general gives better results than its modification (the FIDBFC algorithm).

Table 4. Clustering Accuracy Indices calculated for the fuzzification performed on the Indian Liver Patient Records data set

Algorithm	Partition Coefficient index	Partition Entropy index	Fukuyama-Sugeno index	Xie-Beni index	Purity index	Normalized Mutual Information index
FCM	0,76521	0,51781	7,31219	0,23369	0,71355	0,03833
GK	0,84262	0,33978	4,81232	235,85757	0,71355	0,03798
GG	0,80658	0,41956	6,89524	5,91104	0,71355	0,03553
MID	0,94466	0,12144	7,53296	150,68521	0,71520	0,04287
FEBFC	0,88244	0,25931	3,02895	0,06235	0,71355	0,02400
FIDBFC	0,87023	0,28597	3,33581	0,06514	0,71355	0,02874

As we can see in Table 4, the MID algorithm gives the most accurate clustering results according to the Partition Coefficient, the Partition Entropy, the Purity and the Normalized Mutual Information indices. According to the Fukuyama-Sugeno as well as the Xie-Beni indices, clustering results of the FEBFC algorithm are the most accurate.

Table 5. Clustering Accuracy Indices calculated for the fuzzification performed on the Chronic Kidney Disease data set

Algorithm	Partition Coefficient index	Partition Entropy index	Fukuyama-Sugeno index	Xie-Beni index	Purity index	Normalized Mutual Information index
FCM	0,77893	0,49223	5,33962	0,22215	0,74692	0,17449
GK	0,86890	0,28460	2,22786	169,93521	0,75999	0,19142
GG	0,79741	0,44822	4,80674	22,14820	0,68667	0,13921
MID	0,93745	0,13790	5,55817	301,79223	0,63449	0,07141
FEBFC	0,87330	0,28549	1,50738	0,06734	0,65568	0,09795
FIDBFC	0,82725	0,38268	3,01684	0,08254	0,67831	0,12089

The FIDBFC algorithm on the Indian Liver Patient Records data set is more accurate than the original FEBFC algorithm according to the Normalized Mutual Information index, but less accurate or has the same accuracy according to the other indices. Thus, for the Indian Liver Patient Records data set the FEBFC algorithm in general gives better results than its modification (the FIDBFC algorithm).

As we can see in Table 5, the MID algorithm gives the most accurate clustering results according to the Partition Coefficient and the Partition Entropy indices; the FEBFC algorithm gives the most accurate clustering results according to the Fukuyama-Sugeno and the Xie-Beni indices; the GK algorithm gives the most accurate clustering results according to the Purity and the Normalized Mutual Information indices.

The FIDBFC algorithm is more accurate than the original FEBFC algorithm according to the Purity and the Normalized Mutual Information indices, but less accurate according to the other indices. Thus, for the Chronic Kidney Disease data set the FEBFC algorithm in general gives better results than its modification (the FIDBFC algorithm).

V. Conclusion

In this paper a modification of the FEBFC clustering algorithm was introduced. It is based on using a fuzzy information density instead of the fuzzy entropy measure. The modification, called a Fuzzy Information Density Based Fuzzy Classifier, with several other clustering algorithms were implemented in special software tool, that allowed to make comparison of clustering results. It was experimentally approved, that on some medical data set the introduced FIDBFC algorithm gives better results than the original FEBFC algorithm. Thus, for such data set using the FIDBFC algorithm for transformation from numeric into linguistic and fuzzy values is preferable.

References

[1] Zaitseva E., Levashenko V., Kvassay M., Barach P. *Healthcare system reliability analysis addressing uncertain and ambiguous data.* 2017 International Conference on Information and Digital Technologies (IDT), Žilina, 2017, pp. 442-451.

[2] Barach P., Levashenko V., Zaitseva E. *New Methods for Healthcare System Evaluation Using Human Reliability Analysis.* Proceedings of the Human Factors and Ergonomics Society, vol. 61(1), 2017, pp. 583-587.

[3] Lee H. M., Chen Ch. M., Chen J. M., Jou Y. L. *An efficient fuzzy classifier with feature selection based on fuzzy entropy.* In IEEE Transactions on systems, Man, and Cybernetics – Part B: Cybernetics, vol. 31, no. 3, 2001, pp. 426-432.

[4] Popel D. V. *From continuous to multiple-valued data.* In Proceedings IEEE International Symposium on Multiple-Valued Logic, 2003, pp. 367-372.

[5] Bezdek J. C. *Pattern recognition with fuzzy objective function algorithms.* New York, NY: Plenum Press, 1981, 272 pages.

[6] Dunn J. *A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters.* Journal of Cybernetics, 3 (3), 1974, pp. 32-57.

[7] Gustafson D. E., Kessel W. C. *Fuzzy clustering with a fuzzy covariance matrix.* Proc. IEEE CDC, 1979, pp. 761-766.

[8] Gath I., Geva A. B. *Unsupervised optimal fuzzy clustering.* IEEE Trans. Patt. Anal. Machine Intell., vol. 11, no 7, 1989, pp. 773-780.

[9] Cui H., Zhang K., Fang Y., Sobolevsky S., Ratti C., Horn B. K. P. *A Clustering Validity Index Based on Pairing Frequency.* In IEEE Access, vol. 5, 2017, pp. 24884-24894.

[10] Rendón E., Abundez I., Arizmendi A., Quiroz E. M. *Internal versus external cluster validation indexes.* International Journal of Computers and Communications, vol. 5(1), 2011, pp. 27-34.

[11] Abonyi J., Feil B. *Cluster Analysis for Data Mining and System Identification.* Birkhäuser Verlag AG, 2007, 319 pages.

[12] Rezaee M. R., Lelieveldt B. P. F., Reiber J. H. C. *A new cluster validity index for the fuzzy c-mean.* Pattern Recognition Letters, vol. 19, 1998, pp. 237–246.

[13] Bezdek J. C. *Cluster validity with fuzzy sets.* Journal of Cybernetics, vol. 3(3), 1973, pp. 58-72.

[14] Bezdek J. C. *Mathematical models for systematics and taxonomy*. In: Proceedings 8th International Conference in Numerical Taxonomy, Freeman, San Francisco, 1975, pp. 143-166.
[15] Fukuyama Y., Sugeno M. *A new method of choosing the number of clusters for the fuzzy C-means method.* In Proc. 5th Fuzzy Syst. Symp., 1989, pp. 247-250.
[16] Xie X. L., Beni G. *A validity measure for fuzzy clustering.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 8, Aug. 1991, pp. 841–847.
[17] Zhou K., Ding S., Fu C., Yang S. *Comparison and weighted summation type of fuzzy cluster validity indices.* Int J Comput Commun Control, vol. 9, 2014, pp. 370-378.
[18] Pakhira M. K., Bandyopadhyay S., Maulik U. *Validity Index for Crisp and Fuzzy Clusters.* In. Pattern Recognition, vol. 37(3), 2004, pp. 487-501.
[19] Sripada S. Ch., Rao M. S. *Comparison of purity and entropy of K-means Clustering and Fuzzy C means Clustering.* Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2, No. 3, 2011, pp. 343-346.
[20] Kvålseth T. O. *On Normalized Mutual Information: Measure Derivations and Properties.* Entropy, vol. 19(11), 2017, 631 pages.
[21] *The Home of Data Science & Machine Learning*, https://www.kaggle.com.
[22] *UCI Machine Learning Repository: Center for Machine Learning and Intelligent Systems*, https://archive.ics.uci.edu/ml/index.php.

Satellite Imagery Resolution Enhancement for Urban Area Thermal Micromapping

Iryna Piestova, Mykola Lubskyi, Mykhailo Svideniuk, Stanislav Golubov, Peter Sedlacek

***Abstract*—**The approach to urban area thermal micromapping is presented in this article. The temperature field calculation is performed using the higher spatial resolution thermal emissivity. Logical redistribution is used to equalize the multispectral satellite image bands spatial resolution. Comparison of the output temperature map and the long-wave infrared radiance distribution image is demonstrated.

***Keywords*—** infrared satellite imagery, resolution enhancement, thermal micromapping, urban area.

I. Introduction

In recent decades, there has been an intensive expansion of urbanized landscapes. Industrial production enlargement and population migration from villages to large cities prompt the residential neighborhoods' development. At the same time, the urban landscapes create a specific microclimate, which is characterized by a higher air temperature due to the presence of large amount of areas that capable to accumulation the heat (asphalts, concrete, bitumin etc.). Low concentration of green plantations and significant air pollution, which also contributes the heat retention [1]. Thus, the space monitoring of the urban environment in the longwave infrared band of the electromagnetic spectrum (wavelength of 8-14 microns) is an effective approach for analyzing of the heat load of urban areas to the environment [2].

High heterogeneity is a characteristic feature of urbanized landscapes. Large several of objects and covers with different thermophysical properties are present here. Today, the Landsat-8 is the only one operating satellite, which supplies remote sensing data that allow heat fields mapping with a moderate spatial resolution. It provides data in two long-wave infrared bands (10.3-11.3 μm and 11.5-12.5 μm wavelengths). But, spatial resolution of 100 m is insufficient for a detailed study of the vegetation influence on the urban environment.

However, there are approaches to the use of the visible and near-infrared data, in order to increase the informative and accuracy of the terrestrial objects and covers temperature spatial distribution determining.

II. Concept and Methods

The proposed approach is based on the combined use of thermal imagery by Landsat-8 TIRS sensor (https://landsat.usgs.gov/) and multispectral imagery by Sentinel-2 MSI sensor (https://earth.esa.int/web/sentinel/missions/sentinel-2). Spatial resolution enhancement of the land surface temperature distribution is achieved using the higher spatial resolution thermal emissivity distribution [3]. The land surface thermal emissivities of various pavements are calculated through the data of multispectral imagery with spatial resolution equalization [4].

I. Piestova, Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science, National Academy of Sciences of Ukraine, Kiev, Ukraine (e-mail: pestovai@ukr.net).
M. Lubskyi, Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science, National Academy of Sciences of Ukraine, Kiev, Ukraine (e-mail: N.Lubsky@nas.gov.ua).
M. Svideniuk, Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science, National Academy of Sciences of Ukraine, Kiev, Ukraine (e-mail: m.shved@ukr.net).
S. Golubov, Scientific Centre for Aerospace Research of the Earth, Institute of Geological Science, National Academy of Sciences of Ukraine, Kiev, Ukraine (e-mail: asdfieldspec3@gmail.com).
P. Sedlacek, University of Žilina, Žilina, Slovak Republic (e-mail:peter.sedlacek@fri.uniza.sk).

A. Land surface temperature calculation

The end-product radiance data processing is the land surface temperature using inverse Planck's equation [5]:

$$T = \frac{c_2}{\lambda \ln\left(\frac{\varepsilon(\lambda)c_1}{\lambda^5 L_S}+1\right)} \tag{1}$$

where L_s is the spectral radiance of land surface, $\varepsilon(\lambda)$ is the land surface spectral emissivity, c_1 and c_2 are the first and second radiation constants, λ is the thermal radiation's wavelength in band of sensor sensitivity, T is the land surface physical temperature.

Land surface spectral radiance is gained from the input satellite data by specific procedure of atmospheric correction that allow removing atmospheric influence on radiation [6]:

$$L_S = \frac{L_i - L_i^{\uparrow}}{\varepsilon_i \tau_i} - \frac{1-\varepsilon_i}{\varepsilon_i} L_i^{\downarrow} \tag{2}$$

where L_i is the top of atmosphere (TOA) radiance in i-th band; $L_i^{\uparrow}$ is the upwelling sky irradiance in i-th band; $L_i^{\downarrow}$ is the downwelling scattered radiance in i-th band; τ_i is the spectral atmospheric transmittance.

Determination of the land surfaces emissivity distribution can be performed using multispectral visible and near infrared (VNIR) remote sensing imagery.

Surface emissivity is determined by relationship between emissivity and normalized difference vegetation index (NDVI) distribution. The emissivity is a rather inert surface feature. It is possible to involve data obtained with some time interval relatively long-wavelength infrared imagery.

For soils and vegetation, this dependence is established through the projective vegetation cover fraction F, which is defined as follows [7]:

$$F = \left(\frac{NDVI - NDVI_0}{NDVI_\infty - NDVI_0}\right)^2 \tag{3}$$

where $NDVI$ is the value of the vegetation index in the current pixel; $NDVI_0$ is the maximum value of bare soil index; $NDVI_\infty$ is the vegetation index minimum value of the completely vegetated surface.

The emissivity calculation is carried out as follows:

$$\varepsilon_\lambda = \varepsilon_{v\lambda} \cdot F + \varepsilon_{s\lambda}(1-F) + \Delta\varepsilon_\lambda \tag{4}$$

where ε_λ is the land surface emissivity in current pixel; $\varepsilon_{v\lambda}$, $\varepsilon_{s\lambda}$ are the emissivities for a surface completely covered with vegetation and bare soil or another surface, where there is no vegetation relatively; $\Delta\varepsilon_\lambda$ is an amendment that takes into account the irregularity of radiation due to rough surface (the typical value for a rough surface is approximately 0.005).

Relationship for other types of covers is estimated by the regressive dependence between artificial surfaces spectra taken from ASTER Spectral Library (http://speclib.jpl.nasa.gov) and NDVI index outside its range that corresponds to the vegetation cover and open soil. Based on the obtained spectra a point cloud of dependence of the averaged emissivity of each of the typical spectra from NDVI index is expressed through the sensor-registered values in red and

near infrared (NIR) bands. The resulting cloud of points is averaged quasi-optimally over both axes and then the spline-approximation of the dependence is performed through the obtained knot points.

B. Multispectral imagery spatial resolution equalization

Land surface emissivity distribution obtaining is most effective with the VNIR data obtained by Sentinel-2 series satellites of the in combination with the long-wavelength infrared data provided by the Landsat-8 satellite.

The interpolation for Sentinel-2 MSI sensor spectral bands ground sampling distance (GSD) equalizing must be carried out before analysis. The first step is 20 m spectral bands resampling up to 10 m one. Such procedure generates four identical subpixels. It is necessary to reallocate band signals [8] in these subpixels properly to enhance the overall image actual resolution taking into account the surrounding area context. Selected reallocation matrix is a tool for reclassification [9] of current subpixel after analysis. Its class is considered jointly with near neighborhood, and resulting decision is made to save this one or change it to another class from a set of surrounding subpixels. After additional image adjustments are conducted [10], the Sentinel-2 data acquires an unified 10 m spatial resolution, compared to 30 m for Landsat-8.

This interconnection for other types of land covers is estimated by the regressive relationship between the artificial surfaces spectra taken from the ASTER spectral library and the NDVI index beyond its range corresponding to plant cover and open soil. This technique for obtaining the emissivity distribution of is most effective with the visible and NIR data derived by Sentinel-2 satellites.

C. Emissivity spatial resolution enhancement

Besides extracting land surface temperature from radiance data emissivity distribution, obtained using VNIR data, allow to significantly enhance the spatial resolution of resulting spatial distribution of temperature.

Also we can derive further spatial resolution enhancement by subpixel processing of pair of multitemporal emissivity distributions with subpixel displacement one relatively to another. In this case the initial distributions must contain time-invariant data. Because the emissivity is rather inertial feature of land surface, or absolutely unaltered, so the pair of one's distributions with short time difference (up to 2 weeks) can be engaged.

Subpixel displacement is calculated by relative function between source images using special software [11]. Also it is necessary to create a noise distribution image joint for both input ones, which characterizes a homogenous surface, where any value deviation can be explained as noise. Then fusion of low-resolution input images into a joint resampled image by interlaced scan over the higher resolution grid, taking into account the subpixel's offset and noise matrix is conducted. Next procedures are: the estimation of the inverse operator matrix, enhanced resolution image restoration, and iterative image reconstruction to eliminate irregularities and suppress noise.

Obtained emissivity distribution of enhanced spatial resolution is included into inversed Planck's equation (1) instead of the low spatial resolution one.

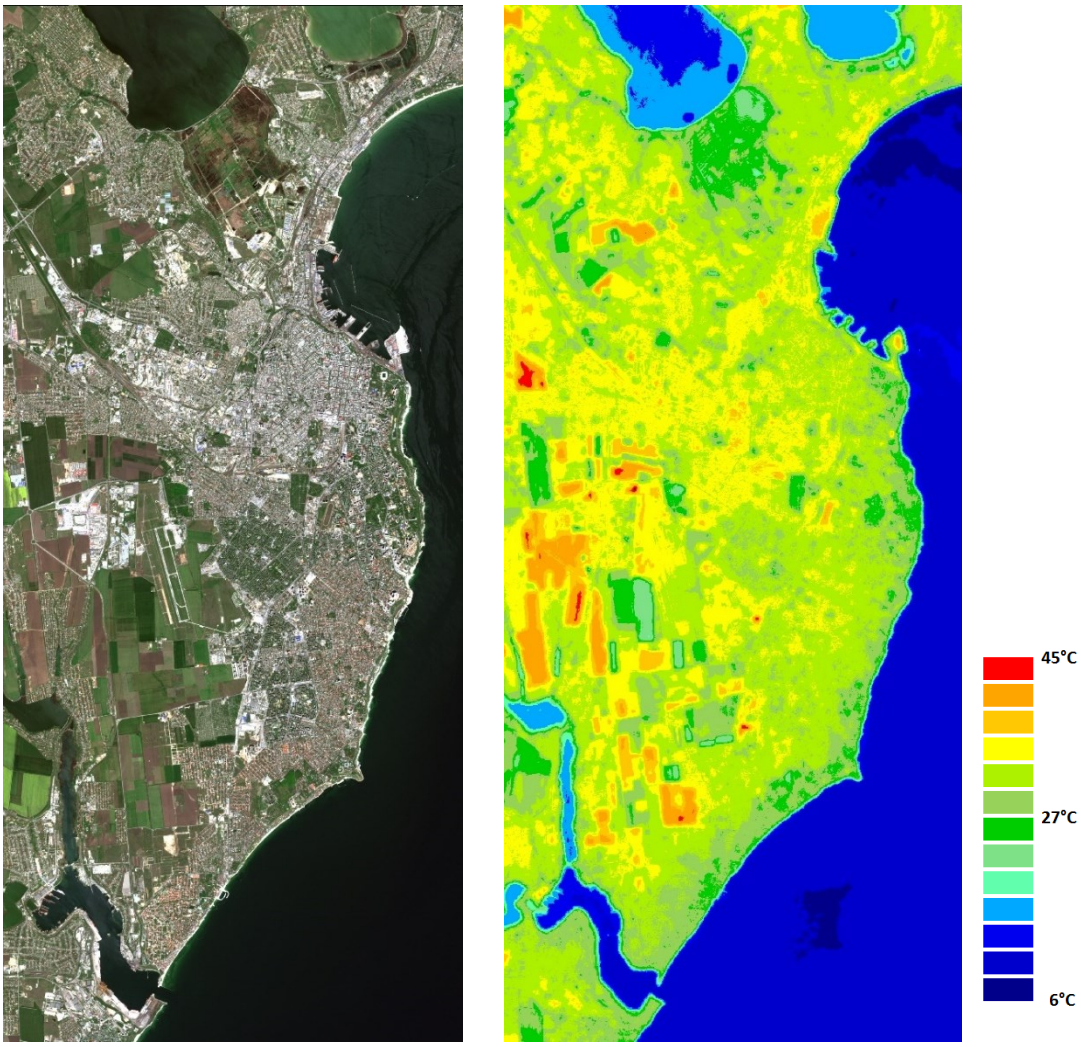
III. Result and Discussion

Odessa city territory was chosen as the object for presented technique validation. The thermal infrared image of Landsat-8 TIRS sensor (23.04.2018, 100 m spatial resolution) and two multispectral images of the Sentinel-2 MSI sensor (28.04.2018 and 30.04.2018, 10 m spatial resolution) were processed.

Multispectral images were used to obtain the distribution of the emissivity. Thermal infrared image was used directly to obtain the enhanced resolution land surface temperature field. The

corresponding software has been developed to perform the procedure for restoring the enhanced resolution image [11].

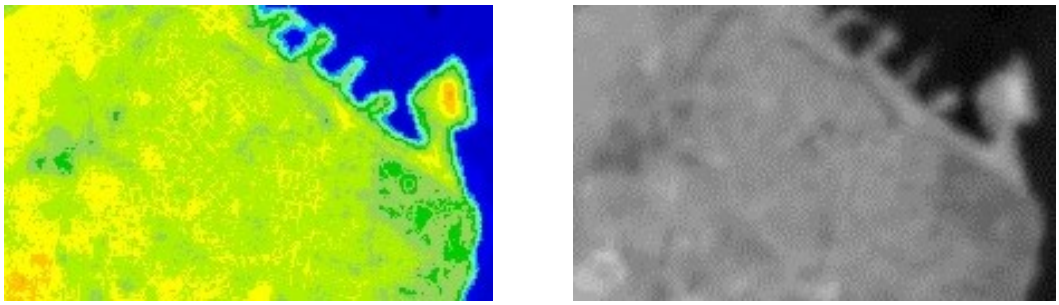
Fig. 1 shows a comparison of the Sentinel-2 satellite multispectral image and the derived temperature distribution map.



a *b*

Fig. 1 Input data and processing result: *a*) Sentinel-2 satellite multispectral image (Odessa city, 28.04.2018, 10 m spatial resolution), *b*) enhanced resolution temperature distribution map

Area of interest covers various land surfaces and a significant building area. Fig. 2 demonstrates a comparison of the output temperature map and the long-wave infrared radiance distribution image obtained by the TIRS sensor.



a *b*

Fig. 2 Thermal field of enhanced detaility (*a*) in comparison with the original TIRS image (*b*)

Enhanced spatial resolution and more fined details are clearly visible. It is also proved that the vegetation areas have a lower temperature than the surrounding buildings and stand out well on the temperature map.

IV. Conclusion

Vegetation mapping with analysis of thermal distribution is a very important for the overall assessment of biomass and its seasonal influence on thermal fields. The moderate resolution satellite imagery was chosen for testing the proposed technique, one of the forthcoming long-term time series analysis.

Subsequent temperature distribution analysis will provide an opportunity to build a time series and to extract the trends of temperature change, as well as the influence of green zones. Which allows to develop a scientific basis for further recommendations to city local authorities and climate change impact mitigation for the urban population.

Acknowledgment

This research was supported by the National Academy of Sciences of Ukraine (grant for research project of young scientists by NAS Presidium decision No. 178 of 14.06.2017) and partly-supported by the National Scholarship Programme of the Slovak Republic for the support of mobility of students, PhD students, university teachers, researchers and artists.

References

[1] C. Yang, X. He, L. Yu, J. Yang, F. Yan, K. Bu, L. Chang and S. Zhang, "The Cooling Effect of Urban Parks and Its Monthly Variations in a Snow Climate City," *Remote Sensing,* vol 9, October 2017, 1066.

[2] S. Wang, Q. Ma, H. Ding and H. Liang, "Detection of urban expansion and land surface temperature change using multi-temporal landsat images," *Resources, Conservation and Recycling,* vol. 128, pp. 526-534, 2018.

[3] S. A. Stankevich, S. V. Shklyar and V. M. Tyagur, "Subpixel resolution satellite imaging technique," in *Proc. Int. Conf. on Information and Digital Technologies (DT'2013)*, Žilina: University of Žilina, 2013, pp.81-84.

[4] I. Piestova, S. Stankevich and J. Kostolny, "Multispectral imagery superresolution with logical reallocation of spectra," in *Proc. Int. Conf. on Information and Digital Technologies*, Žilina: IEEE, 2017, pp.322-326.

[5] H. Tang and Z.-L. Li, *Quantitative Remote Sensing in Thermal Infrared: Theory and Applications*. Berlin: Springer-Verlag, 2014, 281 p.

[6] X. Yu, X. Guo and Z. Wu, "Land surface temperature retrieval from Landsat-8 TIRS – comparison between radiative transfer equation-based method, split window algorithm and single channel method," *Remote Sensing*, 2014. vol. 6(10), pp. 9829–9852.

[7] J. A. Sobrino, J. C. Jiménez-Muñoz and L. Paolini, "Land surface temperature retrieval from LANDSAT TM 5," *Remote Sensing of Environment*, vol.90, 2004, pp.434-440.

[8] M. Kvassay, J. Rabcan and P. Rusnak, "Multiple-valued logic in analysis of critical states of multi-state system," in *Proc. Int. Conf. on Information and Digital Technologies*, Žilina: IEEE, 2017, pp. 212-217.

[9] E. Zaitseva and V. Levashenko, "Reliability analysis of multi-state system with application of multiple-valued logic," *International Journal of Quality and Reliability Management,* vol. 34(6), 2017, pp. 862-878.

[10] S. Stankevich, V. Levashenko and E. Zaitseva, "Multispectral satellite imagery classification using a fuzzy decision tree," *Communications: scientific letters of the University of Žilina,* vol. 16(1), 2014, pp. 109-113.

[11] S. A. Stankevich, S. V. Shklyar, V. N. Podorvan and N. S. Lubskyi, "Thermal infrared imagery informativity enhancement using sub-pixel co-registration," in *Proc. Int. Conf. on Information and Digital Technologies (Rzeszów, Poland, 5-7 July 2016)*, IEEE, 2016, pp. 245-248.

Identification of Medical Three-Dimensional Objects

Tomáš Hlavatý, Lucia Fedoriková

***Abstract*—**This paper deals with identification of objects modelling organs of human body in 3D computer model of the body. For this purpose, we use two existing methods for classification of 3D objects, which are based on finding and comparing features of models of 3D objects. The experiments were done with models of bones of a human body. They showed that the methods are able to classify type of bones, e.g., vertebrae, ribs, bones of arms, but they are not able to recognize a specific bone, e.g., a specific vertebra, rib, or bone of hand or feet. Therefore, we also implemented another method, which tries to take into account layout of the bones in a specific part of the human body. This method is based on measuring a distance between 3D objects. After implementation we experimented with all these methods to find out effectivity of them in various recognition tasks.

***Keywords*—**3D models of human body, classification of 3D objects, descriptor of 3D object

I. INTRODUCTION

Future of majority of sectors is in collaboration with information technology. Thankful technology, we can simulate some processes. For example, one of a European research project founded by the European Union's 7th Framework Program was *RASimAs* project (Regional Anaesthesia Simulator and Assistant). This project aimed at providing a simulator to train physicians in performing regional anaesthesia and an assistant to help anaesthesiologists during the procedure. For this connection of medicine and technology, 3D models of human body are needed. Several such models exist, but they have some issues, which penalize their application [1].

Within RASimAs project a method for rating models was proposed [2]. The method is based on quantification of intersections between the objects that a model of human body is composed of. However, it has some insufficiencies. One of the most important weaknesses of the method is that it recognizes all the intersections as errors. However, some intersections (e.g., between specific vessels and muscles) are natural, and they should not be classified as errors of the model. This implies that intersections should be divided into two groups – "wrong" intersections, which do not exist in a human body and which contributes to the error of the model of a human body, and "good" intersections, which exist in a human body and which should not be taken into account in quantification of the quality of the model. To define, which intersections are good and which are wrong, individual objects of the model have to be classified.

Classification of 3D objects requires the 3D objects to be represented in a way that captures the local and global shape characteristics of the object. This requires creating a 3D descriptor or signature that summarizes the important shape properties of the object. Unfortunately, finding a descriptor that is able to describe the important characteristics of a 3D object is not a trivial task. The descriptor should be able to capture a good balance between the global and local shape properties of the object, so as to allow flexibility in performing different tasks. The global properties of an object capture the overall shape of the object, while the local properties capture the details of the object. Each descriptor has its own strength and weakness for different queries and tasks. According to [3], there are three categories of 3D objects representation: *feature-based methods, graph-based methods*, and *view-based methods.*

Feature-based 3D object descriptors are most popular. They focus on geometric properties of a 3D model to define shape of an object. Some methods like Osada [4] and Ohbuchi [5] include

T. Hlavatý, Faculty of Management Science and Informatics, University of Žilina in Žilina, Slovakia (e-mail: tomas.hlavaty@outlook.com).
L. Fedoriková, Faculty of Management Science and Informatics, University of Žilina in Žilina Slovakia (e-mail: lucka@slavino.sk).

generating points on the surface of the model and measurement relations among them. *Graph-based* methods use the topological information of the 3D object to describe its shape. The graph that is constructed shows how different shape components are linked together. The graph representation includes model graph, Reeb graph [6] and skeleton graphs. These methods are known to be computationally expensive and sensitive to small topological changes. The most effective *view-based* shape descriptor is LightField descriptor developed by Chen [7]. The method captures objects from various angles, so we get set of 2D images.

For the purpose of recognizing objects of a human body, we implemented two feature-based methods. Namely, we implemented Osada's and Ohbuchi's method for creating descriptor. In the experiments, we focused on measuring success of the methods in recognition problem. We have done more experiments like recognition in same model, recognition in various models (Zygote and Anatomium). After getting new model we tried to classify objects. Also we tried to assign a group to an object automatically according another model. At the end, we experimented with a new method thar is based on measuring a distance between objects.

II. Objects Recognition

In this section, we focus on methods for objects recognition that we implemented. At first, we describe existing methods like Osada and Ohbuchi including creation of descriptor and measuring difference between descriptors. The second part is dedicated to definition of our suggested method based on measuring a distance between 3D objects.

A. Osada's Descriptor

Osada's descriptor is creating as follows. Firstly, we generate points at random location on surface of the model. Then distance between every possible pair of generated points ($N(N-1)/2$ pairs for the N generated points) is computed. The descriptor is a 1D histogram created by counting the population of the point-pair distances that fall within a certain distance interval. As it is based on the unoriented point set representation, it is insensitive to the orientation of the surface in the original model. There can be measured different properties, but we have chosen D2 shape function which measure distance between two random points on the surface. This shape function classifies objects better then the other four shape function that were studied in [4].

Generating random points respect surface area of polygonal model. So first we iterate through all polygons and compute its area and store it in an array along with cumulative area of triangles Next, we select a triangle with the probability proportional to its area by generating a random number between 0 and the total cumulative area. To generate a point at random location on the surface of a triangle we use formula $P = (1-\sqrt{r_1})A + \sqrt{r_1}(1-r_2)B + \sqrt{r_1}r_2C$, where A, B, C are vertices of the triangle, and r_1, r_2 are pseudo-random numbers from interval $\langle 0,1 \rangle$ (Figure 1).

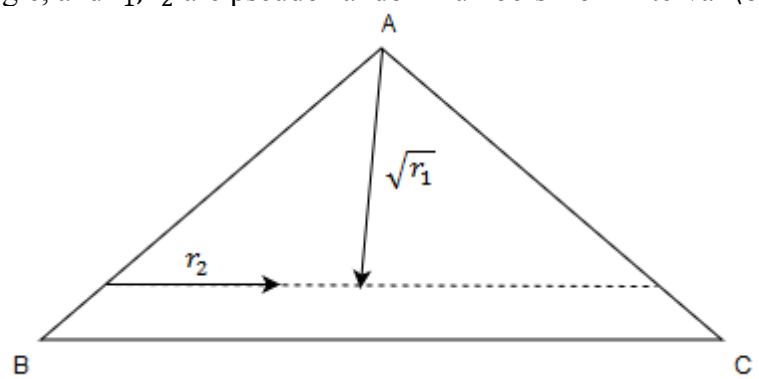


Figure 1 Random point in a triangle

A.1. Comparing Descriptors

Having constructed the shape distribution for two 3D models, we are left with task of comparing them to produce a dissimilarity measure. There are many standard ways of comparing two functions *f* and *g* representing probability distributions [4]. One of the comparing method is the *probability density function* (PDF) norm which outperformed the χ^2 statistics and Bhattacharyya distance and, in general, it is better than the *cumulative distribution function*, because peaks and valleys of PDF curves are easier to discriminate. So, the dissimilarity measure is computed by the mentioned way as follows:

- **PDF L_N:** Minkowski norm of pdf: $D(f,g) = (\sum|f-g|^N)^{\frac{1}{N}}$

Dissimilarity between two descriptors is illustrated in Figure 2, where we can see descriptor of *deltoid* (orange) and *latissimus dorsi* (blue) muscles and highlighted dissimilarity.

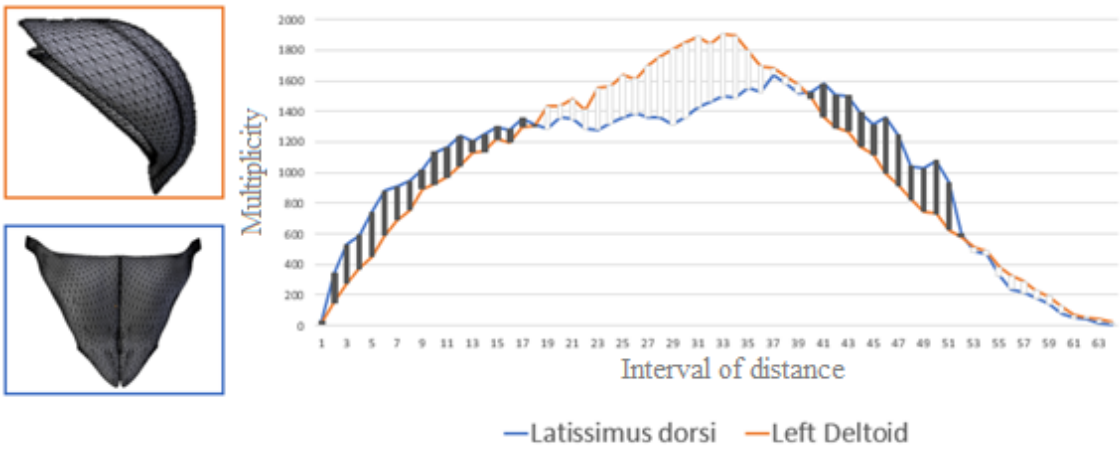


Figure 2 Difference between two descriptors

A.2. Modification of Osada's Method

We add a little modification to the original Osada's method. In this modification we generate *N* pairs of points on the surface of the object. Distance of each generated pair of points is included into histogram. For example, in original method we get 523,776 measured distances with 1,024 generated points. In the modification we need 1,047,552 generated vertices for the same amount of distances. The rest of the steps remains same.

B. Ohbuchi's Descriptor

Another descriptor has been proposed by Ohbuchi [5]. This descriptor is an extension of Osada's descriptor. It uses a Quasi-random number sequence by Sobol [8] for generating numbers r_1 and r_2 for obtaining a point in a triangle of the surface (Figure 1) instead of a pseudo-random number sequence used by Osada. This descriptor also takes into account the inner product of the direction vectors of the generated points.

In this case methods offer two types of descriptors. If the surface of the input model is known to be orientable, we employ *mutual Angle-Distance histogram* (AD) shape feature. If we cannot assume the surface of the models to be properly and consistently oriented, we employ *mutual Absolute-Angle Distance histogram* (AAD). The AD and the ADD shape feature are 2D histograms of distance and angles formed by pairs that are generated on surface of a given 3D object. In computing the AD or the AAD shape feature, an orientation of the point is inherited from the surface normal vectors of polygon on which the points are generated. The angle between a pair of points is represented as the inner product of the direction vectors of the points. The

difference between the AD and the ADD is that the AAD ignores the sign of the inner product. Consequently, the ADD is more robust for models having inconsistent surface orientations than the AD. The AD shape feature measures for each pair of points p_1 and p_2 the 3D Euclidean distance $d = \sqrt{(p_1 - p_2)^2}$ between the points and the inner product $a = n_1 * n_2$ of the direction vectors n_1 and n_2 of the points. Given the distance and the inner product for every pairs of the points, the AD is a joint 2D histogram of the distance *d* and the inner product *a.* ADD histogram is computed like the AD, except that the ADD ignores the sign of the inner product.

B.1. Dissimilarity Measures

Assume that $\mathbf{X} = (x_{i,j})(1 \le i \le \mathrm{I}_d, 1 \le j \le \mathrm{I}_a)$ and $\mathbf{Y} = (y_{i,j})(1 \le i \le \mathrm{I}_d, 1 \le j \le \mathrm{I}_a)$ are the descriptors for models A and B. Descriptors are in fact a 2D matrix of dimension $I_d \times I_a$, in which I_d is a number of distance intervals and I_a is a number of angular (inner product) intervals. The $L1$ norm-based distance $D_{L1}(X, Y)$ and the $L2$ norm-based distance $D_{L2}(X, Y)$ for the AD and AAD descriptors are defined as follows [5]:

$$D_{L1}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{I_d} \sum_{j=1}^{I_a} \left|(x_{i,j} - y_{i,j})\right|,$$

$$D_{L2}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{I_d} \sqrt{\sum_{j=1}^{I_a} (x_{i,j} - y_{i,j})^2}.$$

The $L1$ or $L2$ distance among a pair of column vectors, each of which consists of values from angular bins at the distance bin *i*, is computed first. Then, a simple sum of these distance values over all the I_d intervals is computed

C. Recognition Using Distance between 3D Objects

The second approach to recognition of 3D objects uses distance between 3D objects. To evaluate the similarity of the 3D objects from two models, the general formula for calculation of the Euclidean distance was chosen, i.e.:

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2},$$

where *p* and *q* are points in Euclidean *n*-dimensional space. It is necessary to determine points between which this Euclidean distance will be calculated. So, the point representing the middle point of the object's bounding box displayed as green box in Figure 3 was chosen for this task. These point's coordinates have been obtained by finding the coordinates of two opposing vertices of object's bounding box connected by diagonal and calculating the point lying in the middle of this diagonal. This point was declared as the center of the object's bounding box and therefore as the center of the 3D object. Every such point in 3D space is defined by three coordinates *x, y, z,* so formula for calculating the distance of two such points and thus the distance of two 3D objects in 3D space looks like this:

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

where p_1 and p_2 are points representing the middle points of objects' bounding boxes.

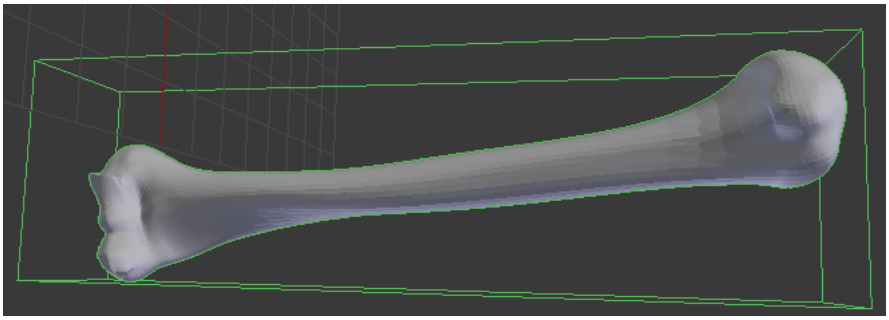


Figure 3 Bounding box of object

The distance between two objects defined by Euclidean distance was used while executing experiments. Object bounded by red box in Figure 4 was an unknown object that we want to recognize. We created capture set of objects bounded by green bounding box in Figure 4 for experimental purposes. They are chosen as random objects in each used anatomical model. In consideration to the objects in such a defined capture set, distance from the object being recognized will be calculated. Measuring distances between unknown objects and objects of this capture set was used as an alternative approach for identifying objects of a human model.

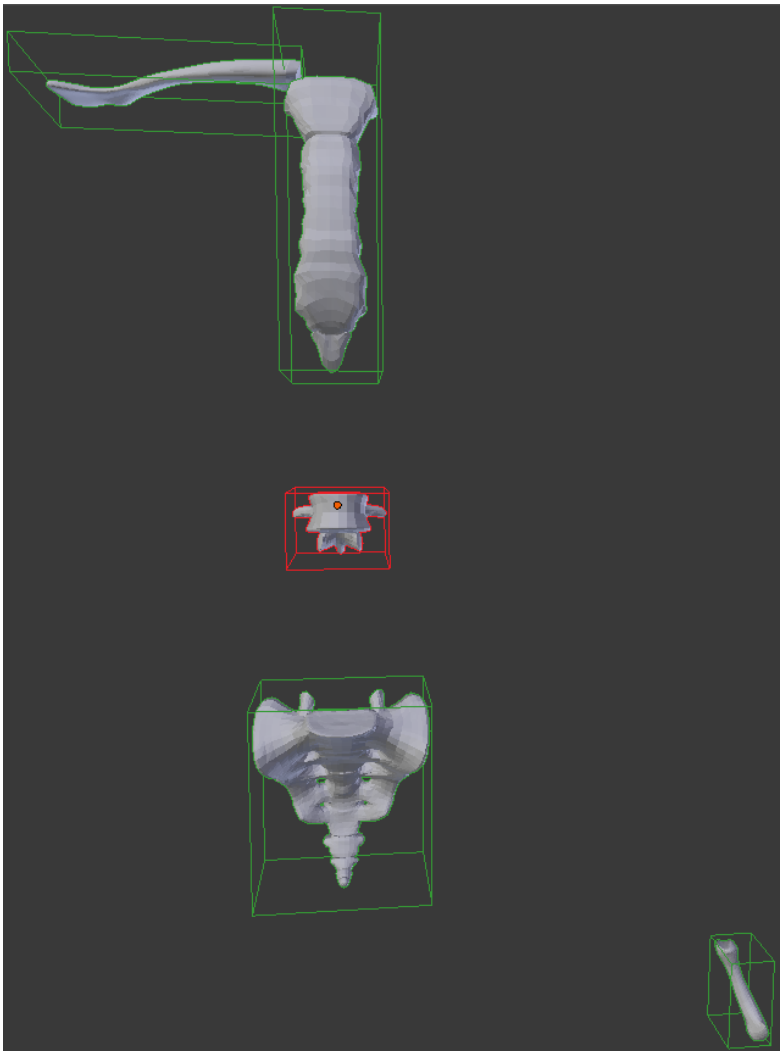


Figure 4 Objects of capture set (green) and searched object (red) with the bounding boxes

III. Experiments

A. Recognizing 3D Objects using Descriptors

The methods described above were implemented as a Blender add-on. In this part, we performed experiments to find out performance of these methods in recognition tasks. We mostly work with two human body models, Zygote and Anatomium. Before the experiments, we tested the connection between amount of generated points and precision. In this test we found out that our modification of Osada's method produce results with better precision, so we used this modification instead of the original method in the experiments. In all the experiments we generated 2,048 random points on surface during creating Ohbuchi's descriptor because it is an appropriate trade-off of performance and time consuming. In Osada's descriptor we generate 1,178,880 pairs of vertices (equivalent of 1,536 generated point). Generating more points did not bring more advantage it just took more time to create the descriptor. Results of all the experiments are presented in Table 1 below.

Table 1 Summary of experiments

Experiment			Skeleton	Muscle
Recognizing object in same model		Osada	149 out of 149	287 out of 287
		Ohbuchi	147 out of 149	270 out of 287
Assigning a group		Osada	87 out of 102	x
		Ohbuchi	89 out of 102	x
Recognizing object in different model	**Top**	Osada	25 out of 103	x
		Ohbuchi	20 out of 103	x
	Top 5	Osada	58 out of 103	x
		Ohbuchi	60 out of 103	x

A.1. Recognizing Object in Same Model

The goal of the first experiment was to recognize and find an object from the first model in the second one only by using a descriptor. At the beginning, we created two same Zygote models of skeleton with left sided objects. We selected one object from the first model and try to look for an object with the most similar descriptor in the second model. If the both object matched, the output was successful. For example, if we selected the skull in the first model, we expected to recognize the skull in the second model. The results of the experiment (the first two rows in Table 1) can be concluded as follows:

- By using modified Osada's descriptor in Zygote skeleton model (dataset in Figure 6) composed of 149 objects (bones), all the objects in the second model were recognized correctly. In Zygote muscle model, which contains 287 left sided objects, 269 objects were correctly identified. Most complicated object to recognize were Rotatores muscles, where one mesh represents more muscles simultaneously, for example Long Rotatores 11 a Long Rotatores 12 were defined by objects with the same mesh. An example of rotatores muscles is in Figure 5.
- By using Ohbuchi's descriptor in the same skeleton model, we correctly recognize 147 objects out of all 149 objects. The first error is observed on vertebra disc and the second is displayed in Figure 7, where the green arrow targets on the selected object and the red arrow on the recognized object. In Zygote muscle model we correctly identified 270 objects out of 287. Most errors are caused by Rotatores muscles as well as we mentioned with Osada's descriptor.

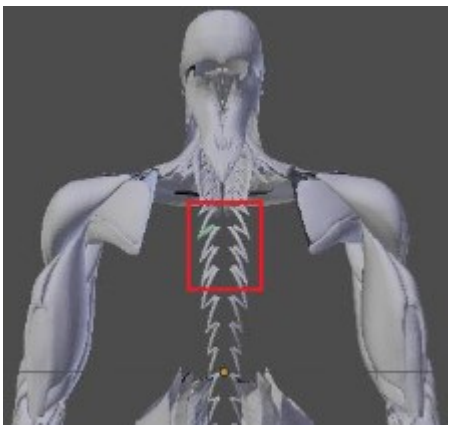


Figure 7 Rotatores muscles

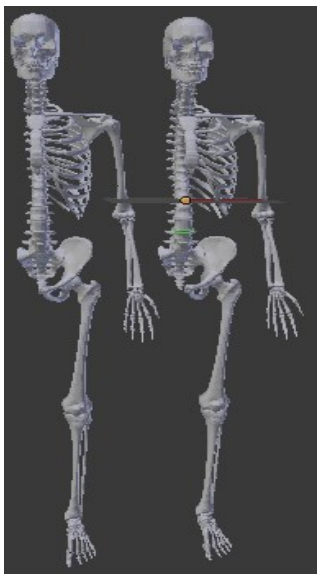


Figure 6 Used dataset

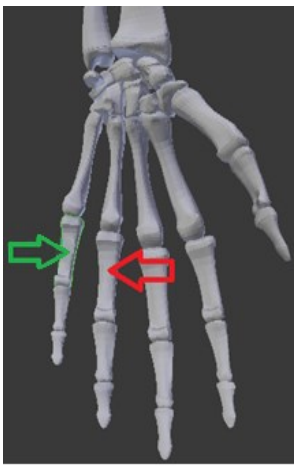


Figure 5
Problematic objects

A.2. Assigning a Group

The second set of experiments was very similar to the first one. Zygote skeletal model was manually divided into five groups (hand, vertebra disc, leg, foot and chest). The Anatomium model was divided at the same way but automatically using a descriptor. We measured how many objects were assigned properly. We obtained the following results (the middle part of Table 1):

- By using modified Osada's descriptor, we achieved that 87 objects out of 102 were assigned to correct group.
- By using Ohbuchi's descriptor, we got 89 objects with a correct group. Most risen errors were caused by bones of hands and feet.

A.3. Recognizing Object in Different Model

In the next set of experiments, we tried to recognize an object in Anatomium model based on descriptors of objects from Zygote skeleton model. So, we selected one object in Zygote model and, subsequently, we try to find this object in Anatomium model. Both models were customized to be compatible, so each of them contains 103 objects. In this case the results were following (the last four rows in Table 1):

- By using modified Osada's descriptor we correctly identified 25 out of 103 objects. In case we would accept correct object occur in top five of the most similar objects instead of the first, we got 58 positives results. The most problematic types of objects were vertebrae because of inequality between vertebrae in Anatomium and Zygote (Figure 9). Other problematic groups were Phalanges Proximalex, Phalanges Mediae a Phalanges Distales which are depicted on Figure 8.
- By using Ohbuchi's descriptor we correctly recognized 20 out of 103 objects. However, if we focus on top five most similar objects we get 60 positive results. Problematic groups of objects are same as in the already mentioned Osada's descriptor.

A.4. Classification of Objects

In this experiment we only focused on problematic objects as vertebrae and bones of hands. In the first part of the experiment, we measured the distance between Zygote objects and Anatomium objects where problem with objects differences occurred. Simultaneously, we obtained a new model [9] so we focused on average distance between pairs of distances Zygote–Anatomium and

Zygote–the new model. If recognized object, i.e., the object with the smallest average distance, matched with a selected object, we classified it as a success.

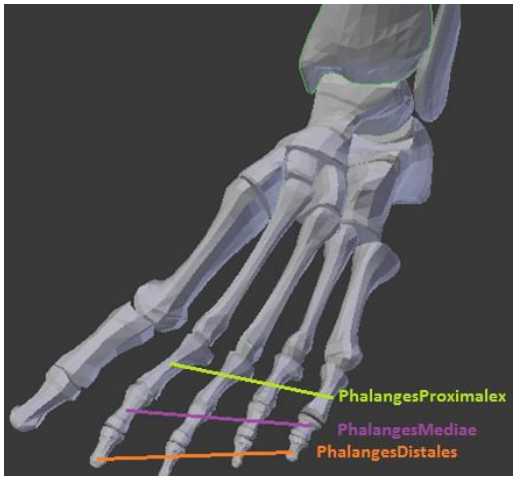


Figure 9 Phalanges Proximalex,Phalanges Mediae and Phalanges Distales

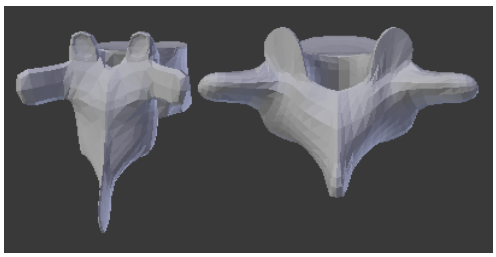


Figure 8 Anatomium ThoracicSpine8 and Zygote ThoracicSpine8

Before applying the new model, we carried out experiments on vertebra column of Zygote and Anatomium. We correctly identified 4 vertebrae out of 25 objects. When we focused on top three most similar objects we got 8 positives output. With added new model, we correctly recognized 6 vertebrae and 13 vertebrae occurred in the top three most similar objects. So, we achieved better results after adding another model. We got all these results with Ohbuchi's descriptor. By using the modified Osada's descriptor, we got the same output as with the previous method. But after adding the new model we correctly identified 9 vertebrae out of 25 object and also 13 vertebrae occurred in the top three most similar objects.

We repeated this experiment with the new dataset which contains bones of hand because this group was also problematic. Ohbuchi's descriptor provided only 1 correctly recognized object out of 19 bones and 10 objects were in the top three most similar object. After adding the new model, we identified 9 object and 13 objects were among the top three most similar objects. By using modified Osada's descriptor we recognized 3 bones out of 19 and 8 objects were in the top tree most similar objects. After adding the new model, we identified 4 objects and 11 objects were among the top three most similar objects. All these results are summarized in Table 2

Table 2 Summary of the second set of experiments

Experiment				**Vertebra bones**	**Palm bones**
Classification objects	Without the new model	**Top**	Osada	4 out of 25	3 out of 19
			Ohbuchi	4 out of 25	1 out of 19
		Top 3	Osada	8 out of 25	8 out of 19
			Ohbuchi	8 out of 25	10 out of 19
	With the new model	**Top**	Osada	9 out of 25	4 out of 19
			Ohbuchi	6 out of 25	9 out of 19
		Top 3	Osada	13 out of 25	11 out of 19
			Ohbuchi	13 out of 25	13 out of 19

B. Identification of 3D Objects using Distance

We assume the correctness of used 3D anatomical models. So, it follows that the relative distances between each two objects should be equal. It is the reason why we decided to use

distance between objects for identification purposes. The results of individual identification methods and performed experiments are summarized in Table 3.

Table 3 Summary of experiments using distance

Experiment		Skeleton	Muscle
Recognizing object in identic model		240 out of 246	501 out of 506
Recognizing object in transformed model		240 out of 246	500 out of 506
Recognizing object in different model	**Legs**	4 out of 11	x
	Vertebral column	2 out of 24	x

B.1. Identification in the Identic Model

Next experiments presented the second approach by using distance between 3D objects. The identic model was created as the copy of the named model, which leave all transformations unchanged. It means both models had the same dimension and rotation. The required element during evaluation of compliance of 3D objects was a capture set of objects containing already named objects, with respect to which the distance of each 3D object is calculated. The match is evaluated when all the distances to the objects in the capture set of objects agree. The success of this recognition method is summarized in the first row of Table 3.

B.2. Identification in the Transformed Model

The same model was used as the transformed model and also as the named model, but with changed dimensions and rotation. Considering the transformation of unnamed model, the match was evaluated in case, when the ratio of objects' volume and the ratio of distance between objects agree. However, the values of ratios varied slightly, therefore the variance of the ratios values was calculated and used as an acceptable variation while the conformity assessment. The match is evaluated if the ratio of objects' volume and the ratio of objects' distance agree. The success of the objects recognition in the transformed model is displayed in the second part of Table 3.

This experiment was also based on distance calculation between objects considering the capture set of objects. However, the difference is in ordering the individual objects according to these distances. We did not consider the summation of all distances among a particular object and objects from the capture set of objects but to the individual distances among these objects. We arranged objects by these distances in ascending order. The match was evaluated when the location of the object in the arrangement matched in both the named and the unnamed model with respect to all objects from the capture object set was same.

As the different model (Anatomium) was used for experimentation and its object were not known and named, it was not possible to evaluate the accuracy of the recognized objects. Therefore, the experiment was performed only on the group of leg bones and vertebral bones. The success of the objects recognition in the different model is summed up in the last part of the Table 3.

IV. Conclusion

We focused on two existing methods and used it for identification of 3D medical objects. We also proposed the method where the distance between 3D objects is used. After implementation of them, we experimented with these methods. We were interested in success of object

recognition in a same model and in various models (Zygote and Anatomium). Subsequently, we got new model where we also tried to classify objects.

We found out that it is needed to generate 2,048 random points on surface during creating Ohbuchi's descriptor because it is an appropriate trade-off of performance and time consuming. In Osada's descriptor we generated 1,178,880 pairs of vertices (equivalent of 1536 generated point). Instead of origin Osada's method we used modified version which generally produce better results than Ohbuchi's one although creating Ohbuchi's one took 50% less time. In the first experiment where we tried to identify objects according a same model we achieved the best results. In the next experiment we tied to automatically assign groups to Anatomium model. Despite of two models we successfully assigned a group to majority of objects. In the third experiment we tried to recognize an object in Anatomium using Zygote objects. In this experiment we got the worst results because Anatomium and Zygote skeletal model are little different, e.g. vertebrae (Figure 9). To improve this output, we added the new model and try to repeat experiments on problematics objects. With this step we reached slightly improvement.

The second approach based on distance between 3D object showed that using distance between 3D objects for object identification purposes is not very successful in case of different models. However, there is some potential for improving the method.

All mentioned approaches have their shortcomings, such as the accuracy of the calculations caused by the accepted variance at compliance assessing. However, if we combined all these methods together, maybe, it could be possible to create a robust approach for identification of objects (organs) in one (unnamed) model based on another (named) model. We plan to solve this problem in our next research. We also plan to combine these methods with other types of methods, such as graph-based or view-based methods.

References

[1] V. Voski, A. Herrler, E. Zaitseva, M. Kvassay, Y. Law, R. B. Duparc, A. Sujar, M. G. Lorenzo, E. Smistad and T. M. Deserno, "A Virtual Physiological Human Model for Regional Anaesthesia," in *Proceedings of Virtual Physiological Human (VPH) Conference*, Amsterdam, The Netherlands, 2016.

[2] E. Zaitseva, M. Kvassay, V. Levashenko, T. M. Deserno, V. Voski and A. Herrler, "Qualitative evaluation of faults (mathematical incorrectness) in anatomical model for Regional Anaesthesia Simulator," in *Proceedings of IEEE Information And Digital Technologies (IDT)*, Rzeszow, Poland, 2016.

[3] I. Atmosukarto, 3D Shape Analysis for Quantification, Classification, and Retrieval, University of Washington, 2010.

[4] R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape distributions," in *ACM Transactions on Graphics*, 2002, pp. 807-832.

[5] R. Ohbuchi, T. Minamitani and T. Takei, "Shape-similarity search of 3D models by using enhanced shape," in *International Journal of Computer Applications in Technology*, 2005.

[6] K. Abhisek, *Introduction to Reeb Graphs.*

[7] D.Chen, X.Tian, Y. Shen and M.Ouhyoung, "On visual similarity based 3D model retrieval.," in *Computer Graphics Forum*, 2003, pp. 223-232.

[8] "http://web.maths.unsw.edu.au/," University of New South Wales, [Online]. Available: http://web.maths.unsw.edu.au/~fkuo/sobol/joe-kuo-notes.pdf.

[9] "http://lifesciencedb.jp/bp3d," lifesciencedb. [Online].

Graphic Tool for Learning and Application of Decision Diagrams in Reliability Analysis

Lukáš Čajka, Martin Belvončík

Abstract—In this article, we present a tool for manipulation with binary decision diagrams. This tool allows applying binary decision diagrams in reliability engineering, i.e., it allows expressing the structure function of the system in the form of a decision diagram and calculate availability of the whole system. The application's role is to make a binary decision diagram based on the structure function, draw it on the screen and calculate availability of the system. On the screen it is possible to move with individual shapes and modify them, e.g., it is possible to change the size, color, and fonts of individual shapes. The created diagram can also be exported to PNG or SVG file format. The standard input for the application is a structure function defined in PLA format.

Keywords—Boolean function, structure function, reliability, binary decision diagram

I. INTRODUCTION

The main objective of reliability as a science is to qualitatively and quantitatively assess the ability of an object to perform and maintain its functions under different conditions. Simultaneously, one of the main goals of reliability analysis is to increase the reliability of the system. For this purpose, it is necessary to [1]:

- find an appropriate representation and model of the system,
- quantify the created model,
- appropriately model and quantify uncertainties that occur in system behavior.

The reliability theory focuses on forecasting, estimating and optimizing product performance. The product can be viewed as a system composed of components. The relationship between the functionality of the components and the functionality of the system is represented by a structure function.

The structure function is used for the mathematical description of the system and describes the system's performance depending on states of its components. For a system composed of n components, this function has n variables that have two possible states (functioning/failure). The structure function then looks like this [2]:

$$\phi(x_1, x_2, \dots, x_n) = \phi(\boldsymbol{x}): \ \{0,1\}^n \rightarrow \{0,1\}, \tag{1}$$

where x_i is a variable representing state of component i, where $i \in \{1,2,\dots,n\}$, and vector $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ is known as a state vector defining states of all the system components.

Structure function defines system topology. However, knowledge of this function is not enough to perform complex reliability analysis. This task also requires knowledge of the state probabilities of the system components. With this information, we can investigate several reliability measures of the system, such as availability (the probability that the system is in state 1) or unavailability (the probability that the system is in state 0), or analyze importance of the system components [3].

It is a problem or a question of how to represent systems that have many components and how to efficiently express their structure functions. One of the possible solutions to this problem is application of approaches of Boolean logic in reliability analysis. This results from a fact that the structure function can be viewed as a Boolean function [4].

L. Čajka, Faculty of Management Science and Informatics, University of Žilina in Žilina, Slovakia (e-mail: cajka.lukas88@gmail.com).
M. Belvončík, Faculty of Management Science and Informatics, University of Žilina in Žilina Slovakia (e-mail: martin.belvo@gmail.com).

II. Representation of Boolean Function

There exist several ways of how to represent a Boolean function: the truth table, logic expression and Binary Decision Diagram (BDD) [5, 6]. Each of them has some pros and cons.

A. Logic Expression

Logic expression consists of logic constants, variables and functions and logic operators (AND, OR, XOR, NOT). The result of the logic expression is one of the values true and false, which are usually represented as numbers 1 and 0 respectively. Examples of logic expressions are conjunctive normal form or disjunctive normal form [7], such as:

$$(\overline{x_1} \wedge x_2 \wedge x_3) \vee (x_1 \wedge \overline{x_2} \wedge x_3) \vee (x_1 \wedge x_2 \wedge \overline{x_3}) \vee (x_1 \wedge x_2 \wedge x_3), \tag{2}$$

where $\wedge$ denotes logical conjunction (AND), $\vee$ agrees with logical disjunction (OR), and $\overline{}$ denotes logical complement (NOT). This expression defines a Boolean function of 3 variables that takes value 1 if not more than one Boolean variable has value 0.

Logic expressions are very convenient for symbolic manipulation, but they are not the best choice for computer implementation and manipulation with Boolean functions because their processing on a computer can be quite complicated.

B. Truth Table

Truth table (Table I) is another typical representation of a Boolean function. This table enumerates all possible combinations of values of the variables of a Boolean function and, for each of them, it contains a corresponding value of the Boolean function. It is very simple for computer implementation and can be processed on a computer very fast. However, its main drawback is a high memory consumption since the size of the table grows exponentially with increasing number of variables (clearly, if a Boolean function has n, then the truth table has to contain 2^n rows). Because of that it can be used just for small functions containing not more than about 30 variables.

Table I Truth table defining the same function as logic expression (2)

x_1	x_2	x_3	$f(\boldsymbol{x})$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

C. Binary Decision Tree

Binary decision tree is a binary tree that has individual variables in the inner nodes, and the decision-making variables in the leaves. Let us suppose we have Boolean variables $x_1, x_2, \dots, x_n$ that make up the input for the function. We test one of the variables in the root, for example x_1 and we have two subtree, one for the case where $x_1 = 0$ and the other where $x_1 = 1$. Each of the two subtrees now tests another variable, each with two other subtrees, etc. On leaves we have either 0 or 1, which are the outputs of the function expressed by a binary decision tree.

An example of a binary decision tree for function defined by logic expression (2) is in Fig. 1. The solid green lines in this figure agree with situations when a decision variable (Boolean variable) takes value 1 and the red dashed lines with situations when the variable has value 0. Every path from the root (it contains variable x_1) to a leaf agrees with one combination of

values of the Boolean variables x_1, x_2, and x_3, and the value in the leaf at the end of the path agrees with the value of the function for this combination.

As one can see, binary decision tree representing a Boolean function of n variables consists of $2^{n+1} - 1$ nodes ($2^n - 1$ inner nodes and 2^n decision-making nodes (leaves)) [5, 6]. This implies its demands on memory are usually greater than demands of the truth table. Because of that, a binary decision tree is not a good data structure for representation of Boolean functions containing a lot of variables. However, it can be reduced in a special form, which is known as a BDD.

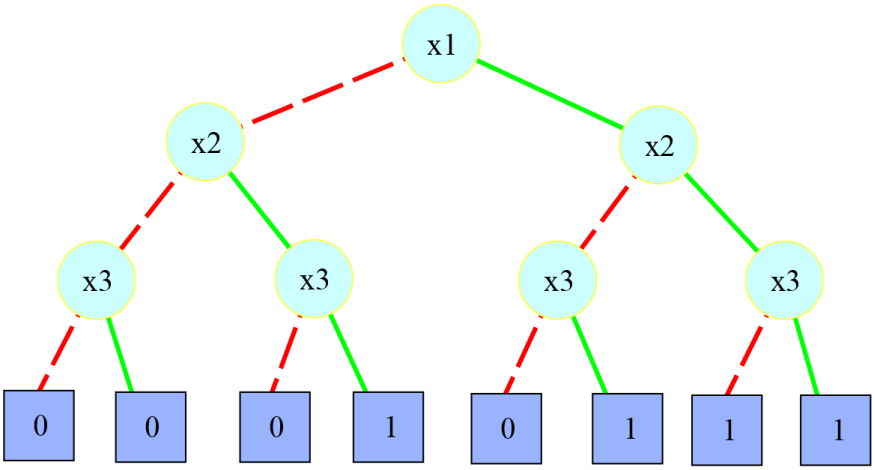


Fig. 1 Binary decision tree defining the same function as logic expression (2)

D. Binary Decision Diagram

BDDs differ from binary decision trees in two directions. Firstly, they allow omitting redundant decision nodes, i.e., nodes that have no influence on the result, e.g., the leftmost node linked with Boolean variable x_3 in Fig. 1. The second improvement is to allow sharing of the same subtrees, e.g., the subtrees defined by two middle nodes linked with Boolean variable x_3 in Fig. 1. This two benefits are achieved by iterative application of two reducing rules, which are known as *elimination of redundant nodes* (Fig. 2) and *merging of isomorphic nodes* (Fig. 3) [5, 6]. For example, if we iteratively apply these two rules on a binary decision tree depicted in Fig. 1, then we obtain a BDD depicted in Fig. 4. As one can see, the diagram is more compact, and it has much less nodes than the tree.

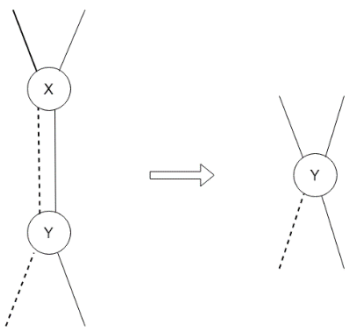


Fig. 2 Rule of elimination of redundant nodes (according to [5])

According to the previous description, a BDD is a quite good choice for representation of Boolean functions because its memory demands are much less than demands of the truth table or binary decision tree, and it can be processed on a computer easier than logic expressions. Thanks to these properties they can be used as an underlying data structure for storing the structure function of systems composed of a huge amount of components. Furthermore, they

can also be used in quantitative reliability analysis because they allow computing system availability and unavailability, which are one of the basic characteristics of systems from reliability point of view, in an efficient way. However, for this purpose, one more information has to be added to the diagram. This information agrees with the state probabilities of the components. In this case, we obtain a new structure, which is known as a probabilistic BDD.

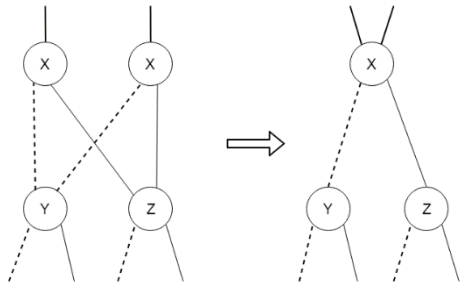


Fig. 3 Rule of merging of isomorphic nodes (according to [5])

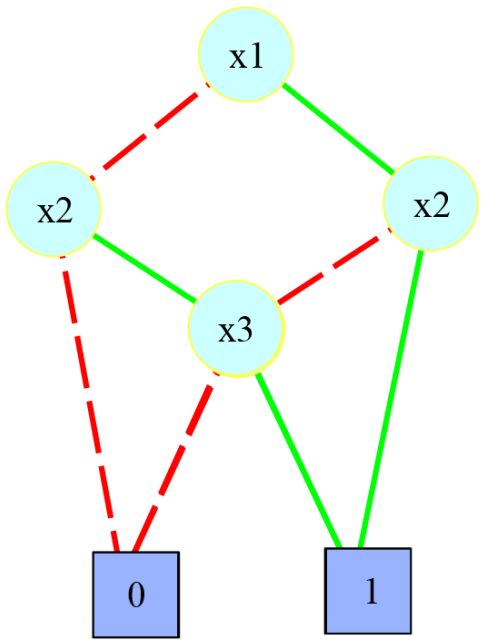


Fig. 4 Binary decision diagram obtained from binary decision tree depicted in Fig. 1

III. Probabilistic Binary Decision Diagram

A BDD is a graphical structure developed for representation of Boolean functions. Since a structure function can be interpreted as a Boolean function, BDD can be used in reliability analysis to represent a structure function. A special type of BDD is a probabilistic BDD [4]. In this kind of BDD, each outgoing edge from an inner node carries additional information that defines the probability that the component (whose behavior is modeled by random variable x_i) associated with the inner node is in a state associated with the outgoing edge.

A probabilistic BDD can be used to calculate the probability of the system state. The probability that the system is in state 0 (does not work) or 1 (works) is simply computable by traversing the BDD and identifying all paths ending in a leaf labeled by number 0 or 1 respectively [4]. This algorithm can be described based on [8] in the following way.

After creating the diagram, for each internal node, we set the probability that a component associated with the node is working. Then we traverse the diagram from the root to leaves and compute the probability that we reach a leaf. At first, we set the probability that we reach the root. This probability agrees with value 1. Then we set the probability of reaching nodes that

are sons of the root. For each node, this probability is set as the product of the probability that we are in the root and the probability that we go from the root to the node. The probability that we go from the root to the specific node agrees with the probability that the component associated with the root is working if we move through the edge associated with state 1 of the component or with the probability that the component is not available if we leave the root by the edge associated with state 0 of the component. When we process all sons of the root, then we move to one of its sons and repeat the previous procedure for this node. After that, we move to another son and repeat the procedure. After processing all nodes at one level, we move to the next level and process all nodes at it. If we visit one node more times (i.e., if a node has more than one parent), we sum all the probabilities of reaching the node. This procedure is repeated until we process all the nodes of the diagram using level order traversal. At the end, the probability that the system is working (available) will be in the leaf denoted by number 1 and the probability that the system is unavailable in the leaf labelled as 0.

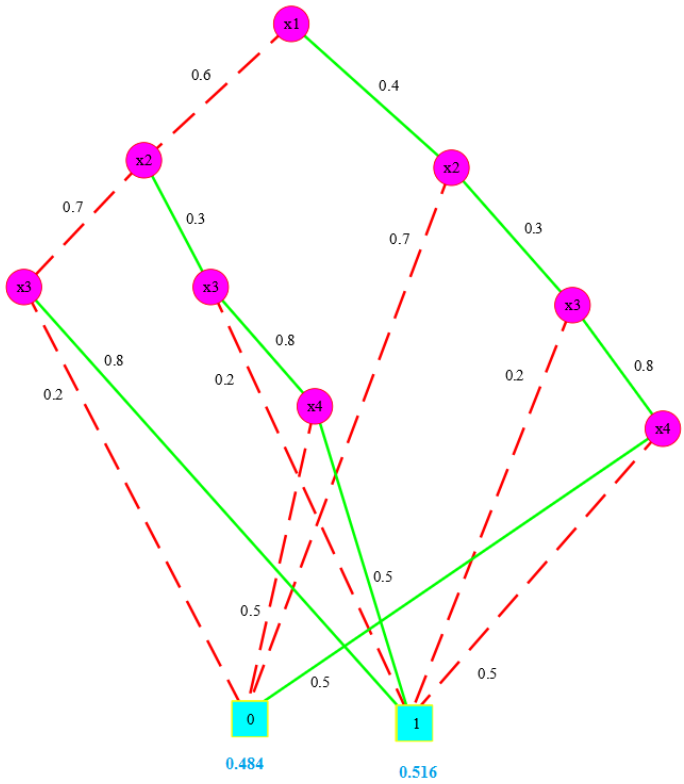


Fig. 5 Example of probabilistic binary decision diagram

IV. Software for Manipulation with Binary Decision Diagrams

BDDs are useful data structure for representation of Boolean functions of high dimensions. They are also useful in reliability analysis of systems composed of a lot of components because they allow representing their structure function in an efficient way. Furthermore, they can also be used to develop efficient algorithms for quantitative reliability analysis. Because of that, they are also taught in the frame of a course on reliability analysis at Faculty of Management Science and Informatics of University of Zilina. However, manipulation with them can be quite complicated without appropriate software support. In a frame of a project course, which is taught in the master study program “informatics” at the faculty, such a software has been prepared in Java. A screenshot of the software can be viewed in Fig. 6.

As you can see, the probabilistic BDD depicted in Fig. 6 is not a real BDD because it contains more than two leaves. However, all other properties of a BDD are satisfied. This form of a BDD can be denoted as an expanded BDD [4]. We choose this special form of BDD because it usually contains fewer crossing lines, what improves its readability.

The software that we created in the frame of the project course allows:

- defining own function,
- automatic creation of a binary decision tree,
- automatic creation of a BDD,
- moving with individual objects on the screen,
- changing specific properties of each object,
- attaching state probabilities to the components linked with internal nodes of a BDD,
- calculating the probability that the system defined by a structure function expressed in a BDD will be working or fail,
- generating the truth table for a BDD,
- loading a Boolean function in PLA file format [9],
- exporting the resulting tree/diagram to PNG or SVG file format.

Thanks to the possibility of exporting the diagram into PNG or SVG file format, it can be used not only for learning of BDDs but also for preparing pictures for teaching materials or for print publications.

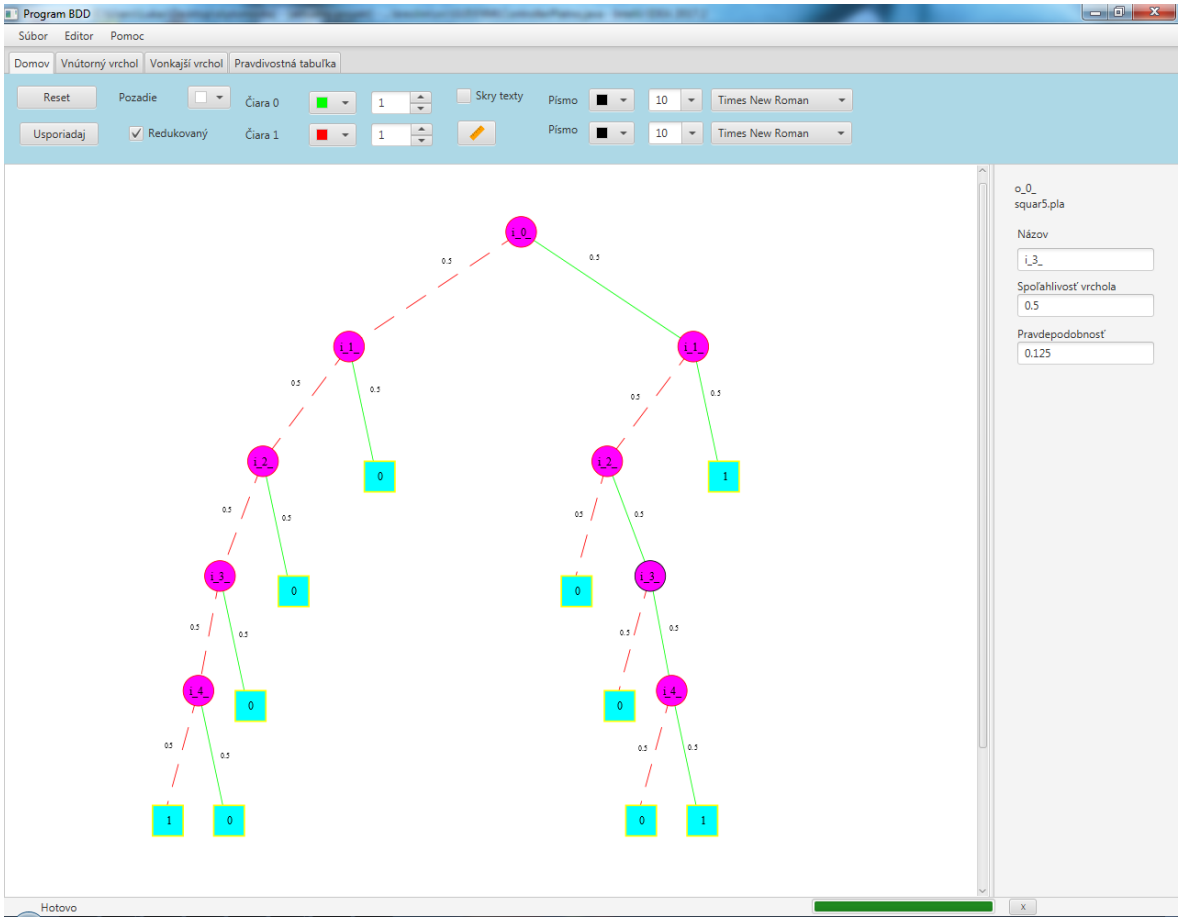


Fig. 6 Software for manipulation with binary decision diagrams

V. Conclusion

BDDs are one of efficient structures for representation of Boolean functions. They can be processed on a computer easier than algebraic formulae, and they have much less demands on memory than the truth table. They can be used in several tasks or research fields, such as logic design or reliability analysis. In reliability analysis, they allow representing the structure function of systems composed of a lot of components. Thanks to this, they can be used to develop new methods for quantitative analysis of complex systems. However, manipulation

with BDDs requires special software because they are not suitable for hand calculations. Because of that, we created a graphic application that allows creating BDDs, manipulating with them and use them in quantification of system availability. The application will be used as a support tool for teaching a course on reliability analysis at Faculty of Management Science and Informatics of University of Zilina.

References

[1] E. Zio, "Reliability engineering: Old problems and new challenges," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 125–141, Feb. 2009.

[2] M. Rausand and A. Høyland, *System Reliability Theory*, 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

[3] W. Kuo and X. Zhu, *Importance Measures in Reliability, Risk, and Optimization: Principles and Applications*. Chichester, UK: Wiley, 2012.

[4] M. Kvassay, E. Zaitseva, V. Levashenko, and J. Kostolny, "Binary decision diagrams in reliability analysis of standard system structures," in *2016 International Conference on Information and Digital Technologies (IDT)*, 2016, pp. 164–172.

[5] T. Sasao and M. Fujita, Eds., *Representations of Discrete Functions*. Norwell, MA: Kluwer Academic Publishers, 1996.

[6] S. Yanushkevich, D. Michael Miller, V. Shmerko, and R. Stankovic, *Decision Diagram Techniques for Micro- and Nanoelectronic Design Handbook*, vol. 2. Boca Raton, FL: CRC Press, 2005.

[7] P. Rusnak, "Transformation of Boolean expression into disjunctive or conjunctive normal form," *Central European Researchers Journal*, vol. 3, no. 1, pp. 43–49, 2017.

[8] J. Kostolny, E. Zaitseva, P. Rusnak, and M. Kvassay, "Application of multiple-valued logic in importance analysis of k-out-of-n multi-state sSystems," in *2018 IEEE 48th International Symposium on Multiple-Valued Logic (ISMVL)*, 2018, pp. 19–24.

[9] "ESPRESSO(5OCTTOOLS)." [Online]. Available: http://www.ecs.umass.edu/ece/labs/vlsicad/ece667/links/espresso.5.html. [Accessed: 12-Jun-2018].

Development of the iKariera Work Portal

Filip Boros, Olga Chovancova

***Abstract*—**The aim of the thesis is the additional development of work portal iKariéra.cz built on Grails technology. The portal should serve students and graduates to find work positions. The application supports language mutations to Slovak and English languages. For the project was done a required upgrade from older to a later version because there is a major difference in both versions. There have been implemented new features for the portal, which were accustomed to requirements of IAESTE Slovakia. Application with the final solution of this work has been deployed to the server. Results of this thesis advanced project in its development and brought it closer to its final completion.

***Keywords*—** Gradle, Grails, Groovy, web application.

I. Introduction

The iKariera work portal will primarily serve students and graduates to find suitable work positions. It should be able to compete with other popular portals with a similar aim. That is why iKariera needs to be easy and comfortable to use and to be built on modern and reliable technologies. The original version of the project was at development phase, which was not yet suitable for deployment. It was built on older versions of technologies. It could not only mean potential security risks, but also lack of performance and stability compared to actual trend. Therefore, the first task was about analysis and upgrading of base technologies of the project. Next work was devoted to the implementation of new features for the portal according to the ideas of IAESTE Slovakia (The International Association for the Exchange of Students for Technical Experience). The project was then deployed to the accessible server for final testing and demonstration of new technologies and features.

II. Used technologies

Portal iKariera is complex web application built on framework Grails. The original version of the project used Grails 2.5.4. The goal was to make the upgrade to new version 3.1.9, which is a complete rework of previous versions. Grails uses and combines advantages of many technologies. It is necessary to mention the most important ones to grasp what Grails is about.

Groovy is an object-oriented programming language for the Java platform. It is very similar to Java programming language. The majority of syntax is similar, but the Groovy aims to simplify specific approaches for Java programmers. Groovy augments and enhances existing Java tools, adds new capabilities to existing Java classes, simplifies tests and many more [1]. Grails uses Groovy as its base programming language.

Gant is a tool for scripting Ant tasks using Groovy. It is used in building processes by Groovy and Java programs. Gant forms integrated a base system for building projects in Grails applications of 1.x and 2.x versions.

Gradle is an advanced tool for automatization of the project building. It combines the best from various build tools and adds its innovations on top of them [2]. Gradle was added for 3.x Grails versions where it replaced Gant. It is used for processes like a compilation, running tests and compression of application. Running Grails command runs competent Gradle command. Introduction of Gradle meant migration configuration files in Grails application into the new *build.gradle* file [3]. This file is also used for the definition of plugins.

F. Boros, University of Zilina, Faculty of Management Science and Informatics, Zilina, Slovakia (e-mail: boros.filip@gmail.com).
O. Chovancova, University of Zilina, Faculty of Management Science and Informatics, Zilina, Slovakia (e-mail: olga@chovancova.sk).

Spring is an open source framework for the development of applications build on Java. It offers complex support for application structure and allows developers to focus more on the development of application itself. Spring consists of many modules which provide different services. It reaches a certain level of abstraction and simplification of code because of that [4]. Grails builds on this framework and uses many of its useful features in processing data and their validation, in configurations, controller logic and many more.

Spring Boot is a tool used for simplification of running Spring-based applications, which otherwise require a lot of various complex configurations. Spring Boot does most of these configurations automatically which saves a lot of time and effort [5]. It became new base upon which are built 3.x versions of Grails.

Grails is web application framework, which contains all aspects of modern web development. It is based on MVC (Model View Controller) architecture. It uses features and technologies from various Java frameworks and combines them with innovations of development with dynamic programming languages. It offers the stability of technologies and simplifies configurations and running applications. It has access to all advantages and features of Java language because it runs on JVM (Java Virtual Machine). At the same time, it is written in Groovy, which allows access to all Groovy libraries and tools [6], [7]. Hibernate is used for the creation of data models and communication with the database. Spring framework is a base controller for Grails application. SiteMesh is a template framework for creating views. Grails binds all of these into one reliable technology.

III. Upgrade of the project

Upgrading from Grails 2.5.4 to Grails 3.1.9 means huge leap in the technological level of the project. It is because Grails 3.x came with most radical changes in Grails development history so far. Versions 3.x are a complete rework of previous versions on a base level. Basic principles remained the same, but the core was rewritten entirely on Spring Boot module and build on Gradle. Many files were removed, changed or added, like new configuration files in format YAML (YAML Ain't Markup Language). These significant changes mean many incompatibilities between both versions [9].

In general, there is need to stick to a sequence of precise steps to achieve fluent and successful upgrade of the project. Main steps are:

- Copy source codes including test files to new Grails 3.x project
- Put plugins definitions into new file *build.gradle*
- Migrate configuration files into new folders
- Remove all unnecessary files

There was also need to upgrade a few plugins into new versions. Plugin *Resources 1.2.7* used for managing static files (JavaScript, CSS) had no upgrade for new Grails version. Therefore it was replaced with plugin *Asset-Pipeline-Grails 2.8.2.*

Another significant change is related to a new feature in Grails 3.x – Interceptors. It is a replacement for filters from earlier versions. They were used to apply specific logic across multiple controllers. In iKariera project were filters used to define default language if none was detected by the user. In upgrade process of iKariera were filters rewritten by new Interceptors which also allowed to use new annotation *@GrailsCompileStatic [8].* This annotation written over method provides static compilations which offer better performance and response of the application.

IV. Implementation of new features

Language mutations to Czech and English languages were present in the original project but were incomplete in some cases and with occasional translation mistakes. Both of these languages were fixed, and there was added completely new translation to the Slovak language. Framework Grails uses for language mutations built-in plugin *i18n.* Translations are stored and defined in separated files. Each for one respective language. They contain expressions with their assigned keys, which are used to call them in the application. Language is switched by changing *Locale* object. Grails provides an easy way to do so by passing a value for parameter *lang* in URL (Uniform Resource Locator). Some pages in the original version of the application were crashing if the user switched languages in them. It was caused by missing URL mappings for these pages. This problem was fixed in this development phase.

Generating CV is important feature implemented for student user account. Generating takes information from the user profile and generates modern CV. The user is shown percentage which reflects completion of his profile. For generating, purposes were in student profile added options for new information such as computer skills, short personal info, and date of birth.

For every new option was added domain model and adjusted competent controllers and views. CVs were implemented in a way that supports translation to currently chosen language in the application. Generating also properly solves the problem with unlisted information from the user profile and keeps the nice final look at CV. For implementing this feature is used iText PDF library. It provides easy creation of a final document by adding phrases and other elements into a document object. The positioning of elements in a document is done by tables. There are created three different templates for generating CV. The user can choose between them in the new view.

New login options are another added feature. At first, it was planned to create a new login with Facebook and Google. However, a recent change in Facebook restrictions requires the application to run on HTTPS (hypertext transfer protocol secure) protocol. The iKariera project in its development phase still runs only on HTTP, so it was decided by IAESTE Slovakia that this feature will be implemented in later development process. Login in with Google remained as the goal for this phase. Log in to an external provider such as Google uses users existing account registered in provider service. After the user clicks on login button, the application sends a request to Google authorization server which will request the user to log in to his Google account.

After successful login, Google service sends special token with user information to the application. Details of this communication can be viewed in Fig. 1. Sharing this discrete information between provider and application uses some of the security protocols. In this case, is used protocol OAuth2. For implementing communication through this protocol is used Grails plugin Spring Security OAuth2.

Some methods of this plugin were rewritten for usage of correct redirects and to customize the way in which is build user account after successful login from Google. If the user logs in with Google and there is found no linked account in an application with Google account, the user is requested to create or link new account in iKariera. This way the user has two separate accounts, one in Google and one in iKariera. This will probably be reduced to just one Google account in future development. However, for now, it is sufficient this way, because it does not interfere with the way how portal account logic works.

New work offers search to filter the last implemented feature. Existing filters in original project are using full-text search provided by SQL (Structured Query Language) to search for single word attributes. This approach is not sufficient for larger text searches in many records. The new filter is searching in job offer description, requirements and other sections for a query defined by the user. For this more complex full-text search approach is used Elasticsearch technology. It is open source tool for full-text search and analysis. It is built on Apache Lucene

which gives it high performance and almost instant search results. It is one of the most popular search engines which offers various features if combined with other tools.

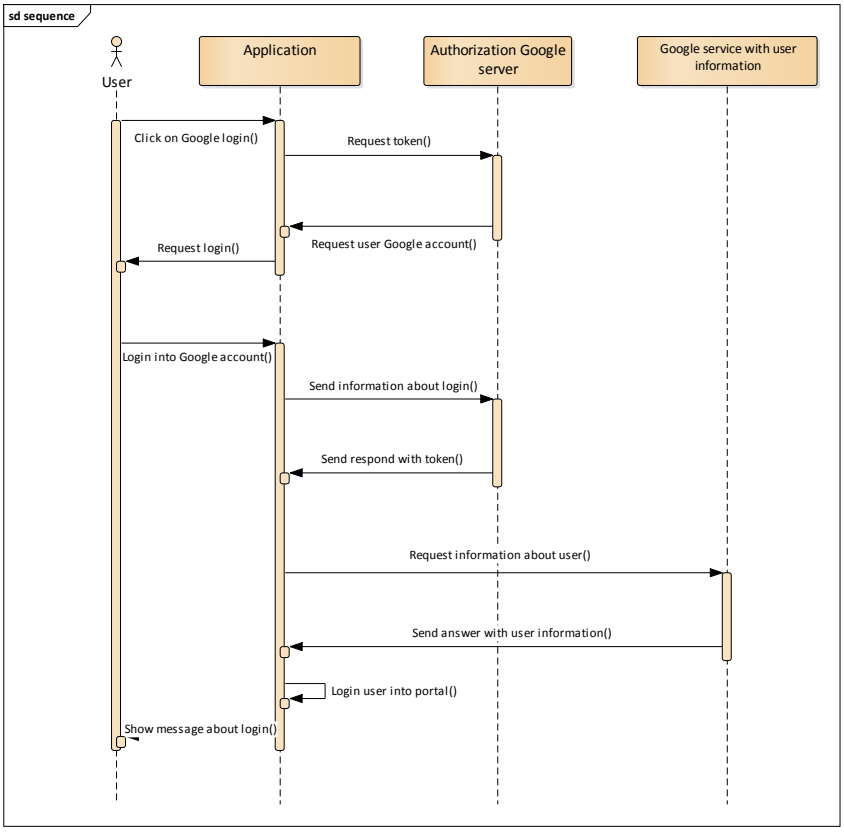


Fig. 1 Sequence diagram of the communication process of Google login

Elasticsearch uses similar principles for storing data in its indices compared to standard databases. It works over a cluster of nodes (server) which enables high-speed searches and protection against loss of data. These are the main reasons for choosing this technology. For implementation is used Grails plugin Elasticsearch 1.2.0. It does abstraction over required configurations on Elasticsearch and simplifies implementation. To enable search over specific records in domain model it is needed to mark them as searchable and give them boost. Boost gives a score to search hits in a specific record. Hits are then ordered by score and returned as a list from Elasticsearch. Other filters then filter the list if the user used them and again ordered by *Top* attribute. Offers marked as *Top* have the highest priority and the second priority is determined by resulting score from search. Example with a table of ordered job offers can be viewed in table 1. Final filtered list of job offers is then displayed to the user.

TABLE I
EXAMPLE OF ORDERED JOB OFFERS AFTER APPLYING SEARCH FILTERS

Order	Job offer	Top	Score
1.	C# developer	Yes	2.05
2.	Java developer	Yes	1.85
3.	Designer	Yes	1.13
4.	Project manager	No	2.36
5.	Service Technician	No	1.03

V. Deployment and testing

The application was after completion of new features deployed to the accessible server for testing purposes. For deployment is used the server provided by IAESTE Slovakia. This server runs on Linux with Ubuntu 14.04.5 distribution. For running application is used Apache Tomcat server. The application requires running instance of Elasticsearch. This needed to be installed first as a service running in the background. Next was application packaged in a WAR (Web Application Resource) format used for static deployment. It was then loaded to Tomcat application folder, and Tomcat server was started. Deployment details of iKariera can be noticed in Fig. 2.

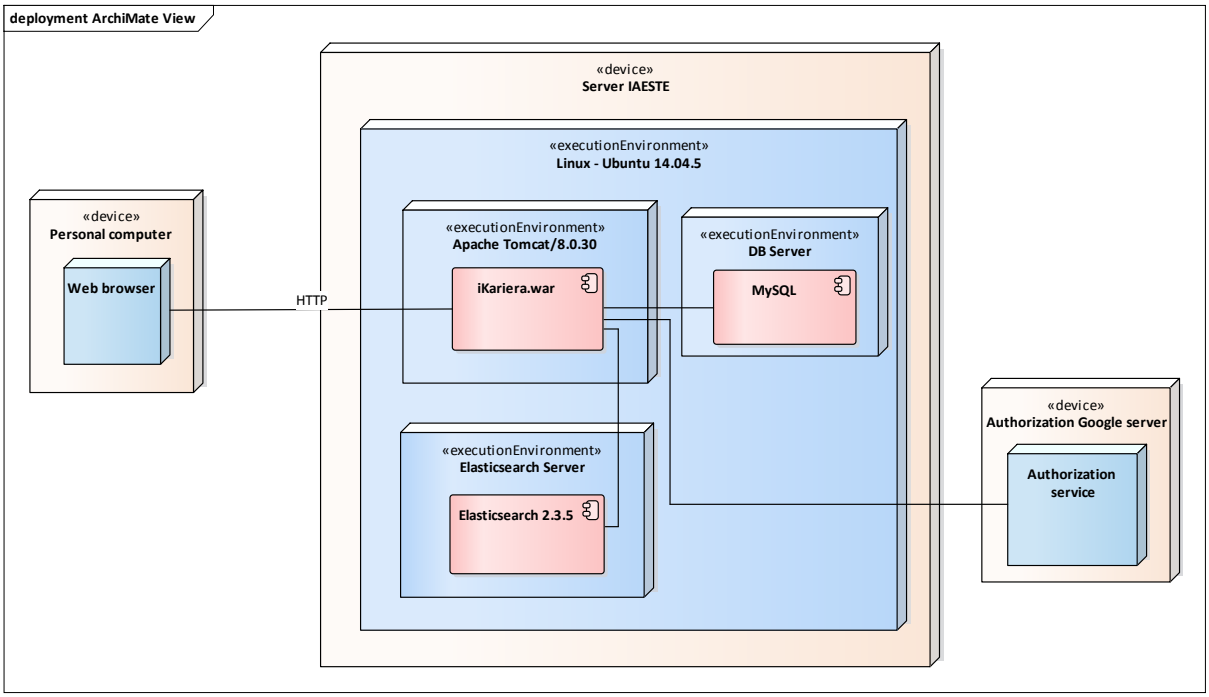


Fig. 2 Deployment diagram of iKariera

Testing of the final version of the application from this development phase was done manually using various versions of all popular web browsers and various devices. For this purpose, was used online tool Browser Sandbox from Turbo.net. Testing was focused on new implemented features and responsiveness of the application. Results showed minor problems with displaying some elements on mobile devices which was subsequently fixed by appropriate CSS rules. There was also a small problem with date input field which was not supported by Internet Explorer and older versions of Mozilla Firefox. Users with these browsers are now allowed to insert date into plain text input field. Overall the application and new features performed well. Performance tests will be used after completion of the application

VI. Conclusion

Portal iKariera is still in its state of developments. It is obvious that many new features still need to be implemented and many features have to be changed. During development of this work were found and fixed various bugs, missing translations and done other minor or more significant fixes. Upgrade of the used technologies proved to be a step in the right direction. Also, the addition of new technologies modernized the whole project and provided new possibilities for the future of the iKariera. The primary aim of this bachelor thesis was to

advance the development of the portal and brought it closer to its completion. It was successfully achieved as the application is now fully functional and accessible for testing purposes.

References

[1] A. K. Kousen, "Making Java Groovy," Manning, pp. 25–47, September 2013
[2] D. K. Abhinandan, "Gradle Essentials," Packt Publishing, pp. 11–30, December 2015
[3] M. Scharhag, "What's new in Grails 3," April 2015, Available at https://www.mscharhag.com/grails/whats-new-in-grails-3
[4] C. Walls, "Spring in Action," Manning, vol. 4, pp. 29–68, November 2014
[5] C. Walls, "Spring Boot in Action," Manning, pp. 12–26, December 2015
[6] C. Olaru, "Grails 3 – Step By Step," Lean Publishing, pp. 10–60, April 2017
[7] G. Smith, P. Ledbrook, "Grails in action," Manning, pp. 16–28, April 2009
[8] P. Behl, "Migrating from Filters to Interceptors," November 2016, Available at http://www.tothenew.com/blog/grails-3-migrating-from-filters-to-interceptors/
[9] M. Ploed, "Grails 3.x update," Video record, Youtube, seen February 2018, available at: https://www.youtube.com/watch?v=IhehO9aM5bk&t=338s