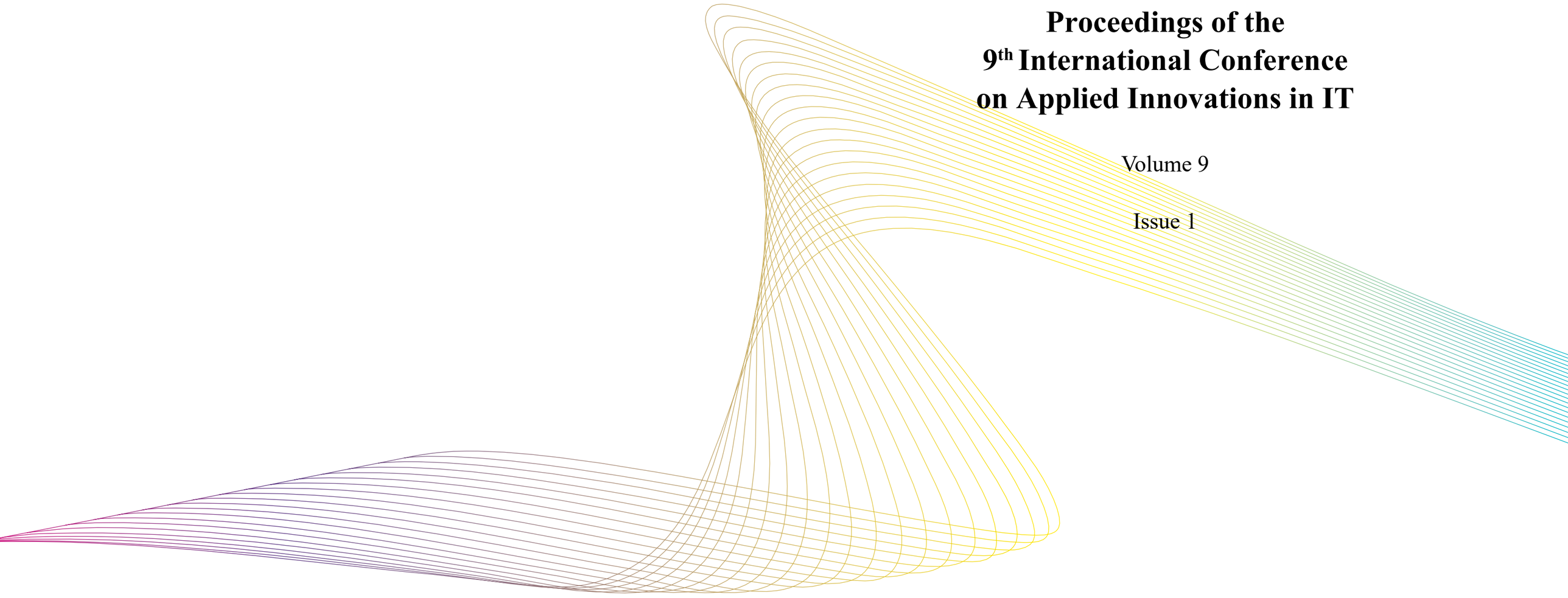
by Anhalt University of Applied Sciences and Perm National Research Polytechnic University

Proceedings of the 9th International Conference on Applied Innovations in IT

Volume 9

Issue 1



EDITION
Hochschule Anhalt

Proceedings of the 9th International Conference on Applied Innovations in IT

Volume 9 | Issue 1

Koethen , Germany
28 April 2021

Editors:

Prof. Dr. Eduard Siemens* (editor in chief),
Dr. Leonid Mylnikov**

(*Anhalt University of Applied Sciences,
** Perm National Research Polytechnic University)

This volume contains publications of the International Conference on Applied Innovations in IT (ICAIIT), which took place in Koethen April 28th 2021. The conference is devoted to problems of applied research in the fields of automation and communication technologies. The research results can be of interest for researchers and development engineers who deal with theoretical base and the application of the knowledge in the respective areas.

ISBN: 978-3-96057-131-5 (Online)
ISSN: 2199-8876

Additional copies of this publication are available from:
FB6 Anhalt University of Applied Sciences
Postfach 1458
D-06354 Koethen, Germany
Phone: +49 3496 67 2327
Email: eduard.siemens@hs-anhalt.de
Web: http://icaiit.org

Content

Section 1. Communication technologies

Section 2. Information Technologies and Cybernetic

FPGA Implementation of IPv6 Header Processor

Zdravko Todorov, Danijela Efnusheva, Ana Cholakoska and Marija Kalendar
Computer Science and Engineering Department, Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Rugjer Boshkovik 18, PO Box 574, 1000 Skopje, N. Macedonia
z_todorov@outlook.com, {danijela, acholak, marijaka}@feit.ukim.edu.mk

Keywords: IPv6 Protocol, FPGA, IP Header Processing, Multi-Gigabit Networks.

Abstract: With the increasing number of Internet devices, the emergence of IoT, 5G and the increased traffic between the devices, the IPv6 is complementing IPv4. As IPv6 is becoming the protocol of choice by the new technologies, in order to accommodate for the features demanded by these technologies it is necessary to have high speed and low latency between the connected nodes. This paper introduces a hardwired IPv6 FPGA node, which processes IPv6 packets and is focused on high-speed transmission. Although, the code is written VHDL, it is written in a way which enables the user to easily add new features and implement new extension headers. The implementation of this IPv6 header processor is done on a Virtex7 VC709 FPGA development board.

1 INTRODUCTION

As of 2021, almost three decades after the appearance of IPv6, only 35% of all accesses to Google have been made with IPv6. Google's chart of accesses starts to grow throughout the year 2011, and the official date of exhaustion of the IPv4 address-space was on 31 January 2011. Even though the adoption rate is not equal in all countries, some countries reach over 50% levels, and some have less than 1% adoption in 2021 [1]. The IPv4 protocol has 2^32 (<4.3e9) possible addresses with total available addresses ~3.7e9, the World's population counts more than 7.5e9, leaving ½ devices for every human. In comparison, IPv6 has 2^128 (~3.4e38) possible addresses (~8e28 more than IPv4). Processing the IPv6 header is different than processing the IPv4. The main difference is the checksum check, which in IPv6 is removed, and instead, bit-level error detection for the entire IPv6 packet is performed by the link layer [2]. Additionally, each device in IPv6 will have its public routable address, which makes it very suitable for the new wireless devices and IoT devices.

The IPv6 processor logic is simple to design, provided that it will be specially adapted to work with IPv6 headers. The proposed header processor will be used to read a single IPv6 header, modify the header where necessary, and then send it to the next node in the network. One of the unique features of the IPv6 protocol compared to IPv4 are the extension headers, which now are of variable non-fixed size, can be placed in mixed order, and only the ones used need to be sent. That means that the protocol by itself requires a certain degree of customizability. Manufacturing such a processor on an ASIC proposes great challenges because of the fast-paced development in the networks and, on the other side, slow-paced IC development (½ to 2 years). Because this type of technology is not suitable for such logic, we are exploring other types of technologies [3].

The FPGA technology suits our requirements with fast-paced development and customizability, and where speed is necessary, we can easily modify the code and trade chip resources for lower latency. In other words, we get a good compromise of performance, price, and re-programmability [4]. This design is developed on a reconfigurable hardware platform – FPGA development board Virtex7 VC709 [5].

The rest of this paper is organized as follows: Section 2 gives an overview of different state of the art solutions. Section 3 describes the proposed IPv6 header processor and explains its ability to provide fast IPv6 header processing. Section 4 presents simulations and synthesis results from the FPGA implementation of the VHDL IP header processor model. Section 5 concludes the paper, outlining the benefits of the proposed IP header processor.

2 STATE OF THE ART

Every network device that is part of a computer network is intended to examine field in the packet headers to decide what to do with each packet. This process of identifying and extracting fields in a packet header is subject to a vast amount of research [6]. With the ever-increasing speed of network links, the research is mostly focused on hardware acceleration for achieving suitable processing speeds [7]. This is mainly achieved by combining application-specific coprocessors with general-purpose multiprocessor systems, or reconfigurable FPGA platforms. In most cases, network processors (NPs) [3] and [8], are used to perform fast data plane packet processing. This includes processing of the IP header, by analysing, parsing and modifying its content. NPs might include some specialized hardware units to perform task offloading, such as lookup and pattern matching, classification of packets, queue management and traffic control [9].

The most popular NPs used today, include one or many parallel homo- or heterogeneous processing cores. For instance, Intel's IXP2800 processor [10], includes 16 identical multi-threaded general-purpose RISC processors organized as a pool of parallel homogenous processing cores that can be easily programmed with great flexibility towards ever-changing services and protocols. Furthermore, EZChip has introduced the first NP with 100 ARM cache-coherent programmable processor cores [11], that is by far the largest 64-bit ARM processor yet announced.

The discussed NPs confirm that most of the operations in NPs are performed by general-purpose RISC-based processing cores as a cheaper but slower solution, combined with custom-tailored hardware that is more expensive but also more energy-efficient and faster. If network packet processing is analysed on general-purpose processing cores, then it can be easily concluded that a significant part of processor cycles will be spent on IP packet header parsing and processing.

On the other hand, some proposals of TCP/IP offload engines [12] provide a certain amount of processing relief compared to a classical network interface card, but still, it requires a huge portion of data processing from the main processor. The sequential software flow, i.e. protocol processing consumes CPU time and resources, creating a dependency between processor load and available throughput as well as latency. This reveals a major drawback, especially for embedded systems where resources are even more limited and CPU time is needed for application-specific tasks. To overcome these system-dependent limitations in throughput and latency, the authors of [13], implement a complete TCP/IP stack in hardware. This 10 GbE hardware-based TCP/IP stack can handle a single physical network interface and contain IPv4, ICMPv4, TCP and UDP protocols.

Furthermore, the authors of [14] introduce a novel architecture implementing a TCP/IP stack capable of processing 10 Gb/s data full-duplex, while handling thousands of concurrent sessions. The architecture's resource requirements scale linearly with the number of supported sessions to over 115,000 given today's 20 nm devices. Similar types of architectures appear in [15] and [16] - the first being an open-source Gigabit Ethernet TCP/IP IPv6 networking architecture, designed for packet processing, IoT, test & measurement, and control (e.g., sensors, motors, etc.) applications and the second implementing a UDP/IP hardware protocol stack that enables high-speed communication over a LAN or a point-to-point connection. The core, designed for standalone operation, is ideal for offloading the host processor from the demanding task of UDP/IP encapsulation and enables media streaming with speeds up to 100Gb/s even in processor-less SoC designs. Assuming that most of the available research presents a hardware implementation of IPv4 protocol, our research is focused on developing a dedicated processor module for IPv6 protocol. The emergence of novel technologies, such as 5G, together with the significant expansion of devices on the Internet and the Internet of Things, makes IPv6 protocol a necessity, compared to IPv4.

3 DESIGN OF IPV6 HEADER PROCESSOR

The IPv6 header processor consists of three processors: main processor, memory processor, and error processor, which is shown in Figure 1. The main processor is processing the header information while receiving data from the memory processor and is sending error data to the error processor. The memory processor is the bridge between the on-board RAM memory and the main processor. Its purpose is to get the necessary data from the external memory and prepare it for the main processor. The error processor reads error data from the main processor and sends the error messages back to the source if necessary.

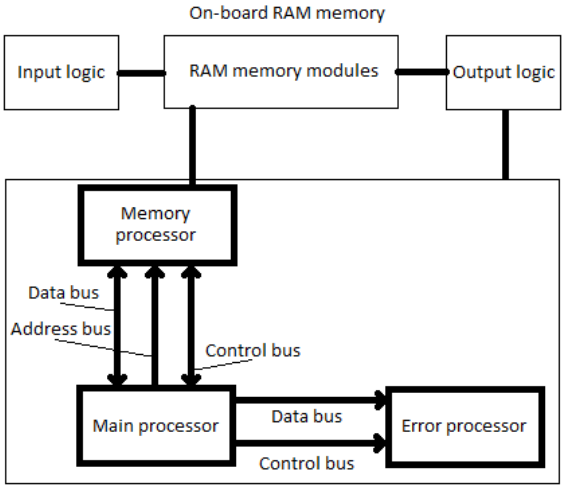


Figure 1: IPv6 header processor internal structure.

This device is designed to work with external on-board RAM. The type of RAM is defined in the memory processor. IPv6 headers have a maximum length of 64kB, meaning that storing the whole header inside the core would make the hardware too complex. The main processor consists of three data buffers, one main 16 octet data buffer and two shared 16 octet data buffers. The communication with the memory processor is realized through a 128-bit data bus and 16-bit address bus. The main data buffer is used for reading the current header continuously. Because there can be more than one main processor in the whole implementation, the shared memory can be reserved for use by each separate main processor.

The size of the main buffer can be changed. The secondary buffers are set to be 16 octets because we can exchange two addresses in a single cycle.

Transferring data from memory to processor takes three cycles with a 128bit data bus and a 16bit address bus. The formula for necessary cycles (NC) to transfer 128bit data given in (1),

$$NC = \lceil 128/N \rceil * 3 \qquad (1)$$

where 1<=N<=128, and N is the width of the data bus in bits. The address bus can be further narrowed. This changes the formula for cycles and adds further complexity to the circuit. For devices with low latency, N=128 is recommended, and for devices with lower logic space, N=16 is recommended. Transferring a whole datagram of 40B takes 9 cycles.

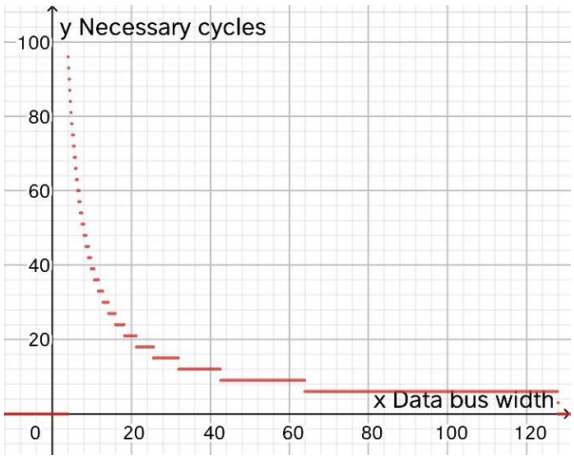


Figure 2: Necessary cycles to transfer 16 data octets to the main processor.

Processing the IPv6 header is done in two phases. The first phase is with the first eight octets of header data, which are always in the same position. Therefore, this data is processed in parallel, and if any errors are detected, the processor sends unique error detection bits to the error processor. If the header needs to be destructed, the memory processor gets this information so that the header can be deleted from the memory.

The first 8 octets of the IPv6 header contain the following information:

- Version - 4 bits are used to indicate the version of IP and is set to 6;
- Traffic class – is available for use by originating nodes and/or forwarding routers to identify and distinguish between different classes or priorities of IPv6 packets;
- Flow label – may be used by a source to label sequences of packets for which it requests special handling by the IPv6 routers, such as non-default quality of service or "real-time" service;
- Payload length – Length of the IPv6 payload, i.e., the rest of the packet following this IPv6 header, in octets;
- Next header – Identifies the type of header immediately following the IPv6 header;
- Hop limit – Decremented by 1 by each node that forwards the packet. The packet is discarded if Hop Limit is decremented to zero;

The following 32 octets contain address information, and once this information is checked then the first phase is finished.

Once the first phase is complete, the main processor starts to process the extension headers in the second phase. The extension headers are continually processed. In this implementation, we are working on a node that is placed between the source and the destination host. Therefore, we are processing only the extension headers which are subject to a change, and we are checking for errors in the fields which need not be changed. The extension headers are processed in the order in which they are present. In RFC 2460 [17], it is recommended that the extension headers are placed in a particular order, but that is not necessary. Additionally, not all extension headers are required to be present. Because of these requirements, the extension headers are processed continually.

Processing the first eight octets of the IPv6 header takes one cycle. Processing of the extension headers depends on whether the header is changed or checked.

In this implementation, we added processing of the routing header extension as an example of the possibilities that this device provides. As an example, once the routing header extension 43 is detected, the processor detects errors and processes the header. The header states that two addresses need to be exchanged and the addresses are stored in the shared memory in order to be exchanged. The shared memory is reserved only for a small portion of time in order to provide possibility of multiple main processors inside a single IC.

Once all extension headers are processed, the second phase is finished. When the second phase finishes, the processor waits for the next IPv6 header. The complete data flow diagram of IPv6 header processing is shown in Figure 3.

The error processor is a separate module which communicates with the memory processor and each main processor. Each main processor can signal the error processor for an error. The error processor then decides whether the packet should be discarded and if so, the error processor sends location information to the memory processor.

The memory processor is a bridge module that connects the main processor with the on-board RAM. When the main processor signals the memory processor for necessary data, the memory processor calculates the location of the data in the RAM. The memory processor has IP (intellectual property) core for communication with onboard DRAM provided by VIVADO Suite. In this way the data is read and written from the DRAM and sent to the main processor.

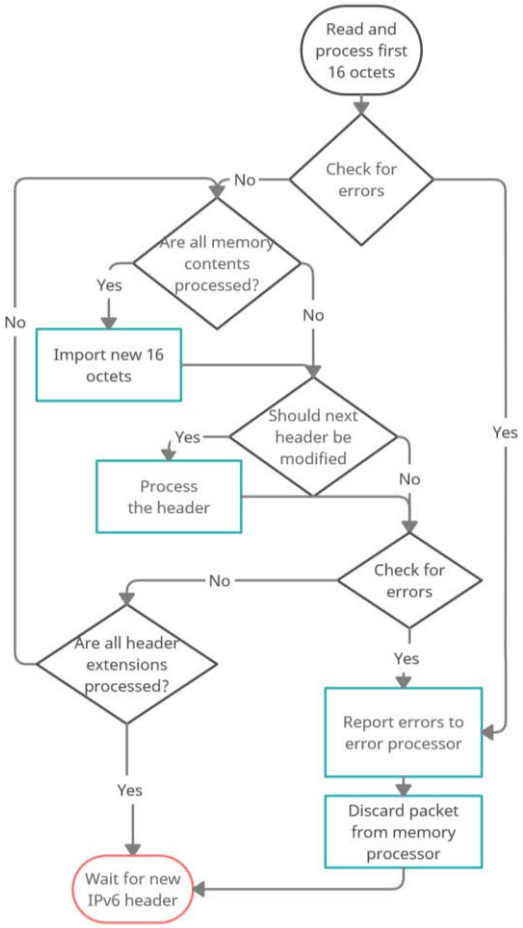


Figure 3: Main processor logic diagram.

4 FPGA IMPLEMENTATION OF IPV6 HEADER PROCESSOR

The proposed IPv6 processor was described in VHDL using the Xilinx VIVADO Design Suite tool. This software environment includes a simulator for performing functional analysis of VHDL models and several other hardware syntheses and FPGA implementation tools.

Simulations and functional analysis were made only for the main processor and the error processor, because the memory processor is implementation-dependant.

Once the analysis is finished, the IPv6 header processor is synthesized and implemented in Virtex7 VC709 evaluation board [5]. The synthesis results

show that the IPv6 header processor can be implemented in the Virtex7 VC709 development board, by utilizing 962 FF and 2653 LUT without the memory processor. More detailed results of the FPGA utilization, after the synthesis of the proposed IP header processor in VC709 FPGA board is shown in Table 1. Furthermore, Figure 4 presents the implemented IP header processor, after the place and route on the appropriate VC709 FPGA board is finished.

Table 1: Utilization of Virtex7 VC709 FPGA resources for the proposed IP header processor.

Resource	Utilization	Available	Utilization %
LUT	2653	433200	0.61
LUTRAM	16	174200	0.01
FF	962	866400	0.11

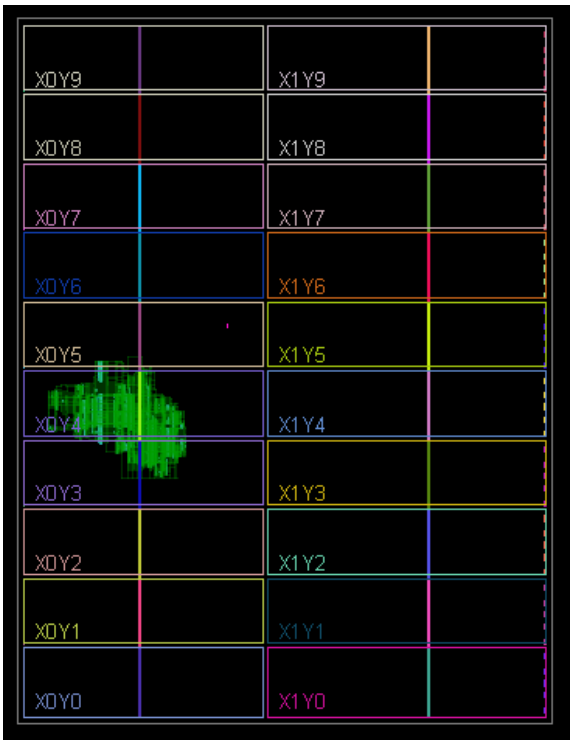


Figure 4: Implemented IP header processor on Virtex7 FPGA board (after place and route).

As a result of the low FPGA resource utilization, the processor can be further extended and then implemented on the same development board. According to this, the FPGA technology makes the proposed processor very flexible and cheap for implementation. Additionally, the ability for ease FPGA reconfiguration, makes the IP header processor implementation suitable for further modifications and improvements.

5 CONCLUSION

The main focus of this paper is the FPGA implementation of the proposed IPv6 header processor. Considering that the implemented IP header processor utilizes less than 0.11% FF and 0.61% LUT FPGA resources, future work would include the whole implementation of the IPv6 processor, including communication with on-board RAM and ethernet port IO. It is evident that these modifications would require more resources than previously used, but this makes the IPv6 packet processor a whole. The possibility of generating various bus widths and different logic will make this kind of processor suitable for less resourceful and powerful FPGA boards.

This device will be very practical in device-to-device communication, because with the implementation of this code in every device, all of the devices will be able to be used as a link between the source and destination node.

This approach makes use of FPGA re-configurability, which has proven to be an ideal solution for achieving reasonable speed at low price.

REFERENCES

[1] Google, IPv6 adoption in the Internet [Online]. Available: https://www.google.com/intl/en/ipv6/statistics.html, 2021.

[2] Ch. M. Kozierok, The TCP/IP Guide: A Comprehensive, Illustrated Internet Protocols Reference, 1st ed. CA: No Starch Press, 2005.

[3] P. C. Lekkas, Network Processors _ Architectures, Protocols and Platforms (Telecom Engineering). McGraw-Hill Professional, 2003.

[4] J. M. P. Cardoso and M. Hubner, Reconfigurable Computing: From FPGAs to Hardware/Software Codesign, NY: Springer-Verlag, 2011.

[5] Xilinx, VC709 Evaluation Board for the Virtex-7 FPGA, User guide, 2016.

[6] G. Gibb, G. Varghese, M. Horowitz, and N. McKeown, “Design principles for packet parsers,” in Proc. of ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 13–24, 2013.

[7] J. Kořenek, “Hardware acceleration in computer networks,” in Proc. of 16th International Symposium on Design and Diagnostics of Electronic Circuits Systems, 2013.

[8] R. Giladi, Network Processors - Architecture, Programming and Implementation, Ben-Gurion

University of the Negev and EZchip Technologies Ltd., 2008.

[9] B. Wheeler, A New Era of Network Processing. LinleyGroup Bob Wheeler's White paper, 2013.

[10] Intel, Intel® IXP2800 and IXP2850 network processors, Product Brief, 2005.

[11] B. Doud, "Accelerating the data plane with the Tilemx manycore processor," in Linley Data Center Conference, 2015.

[12] Z. Bokai, Y. Chengye, and C. Zhonghe, "TCP/IP Offload Engine (TOE) for an SOC System" in Nios II Embedded Processor Design Contest-Outstanding Designs, 2005.

[13] U. Langenbach, A. Berthe, B. Traskov, S. Weide, K. Hofmann, and P. Gregorius, "A 10 GbE TCP/IP Hardware Stack as part of a Protocol Acceleration Platform," in Proc. of 3rd IEEE International Conference on Comsumer Electronics, 2013.

[14] D. Sidler, G. Alonso, M. Blott, K. Karras, K. Vissers, and R. Carley, "Scalable 10 Gbps TCP/IP Stack Architecture forReconfigurable Hardware," in Proc. of 23rd IEEE Annual International Symposium on Field-Programmable Custom Computing Machines, 2015.

[15] Mind Chasers, Private Island: Open Source FPGA-Based Network Processor for Privacy, Security, IoT, and Control, White paper, 2020 [Online]. Available: https://mindchasers.com/education.

[16] Xilinx, UDPIP-100G100G UDP/IP Hardware Protocol Stack, Product Brief, 2020.

[17] Internet Society, "Internet Protocol, Version 6 (IPv6)," RFC 2460, 1998.

Evaluation of the Noise Immunity of the MIAM Communication System

Ivan Pavlov, Valery Lebedyancev, Sergei Abramov, Maria Pavlova and Eugenia Abramova
Federal state budgetary educational institution of higher professional education, Siberian state University of Telecommunications and Informatics, Kirov Str. 86, 630102 Novosibirsk, Russian Federation
IIPavlov02@mail.ru, lebvv1951@mail.ru, abramov@sibsutis.ru, mspavlova@ngs.ru, evgenka_252@mail.ru

Keywords: Group of Transformations, Communication Channel, Invariant Group of Transformations, Modified Invariant Amplitude Modulation, Noise Immunity, Invariant Communication System.

Abstract: The article proposes a modification of the known invariant amplitude modulation that transmits the values of information elements by the ratio of the lengths of the signal vectors lying on a straight line passing through the origin of the coordinate system of the signal space [1]. Modification of this modulation makes it possible to use signals whose vector ends lie on a straight line that does not necessarily pass through the origin of the signal space coordinate system. This gives the opportunity to use in a greater variety of signals, not just signals of similar shape as in the well-known invariant of amplitude modulation that can be useful to enhance immunity against a specific type of interference and to secrecy of messages transmitted. The article contains an assessment of the noise immunity of a communication system with modified invariant amplitude modulation to white noise and a description of its structural scheme.

1 INTRODUCTION

Currently, it is established [1] that the effect on the signals of the communication channel can be reflected by the corresponding group of transformations. For example, the change in signals in linear channels is described by an affine transformation group, a subgroup of which is the group of orthogonal transformations. The latter corresponds to the case when the channel has a flat amplitude-frequency and linear phase-frequency characteristics. The effect of additive interference is displayed by a group of shifts of the ends of the signal vectors in the direction of the interference vector.

The proof of the possibility of describing a channel by a transformation group opens up a method for undistorted message transmission by using group invariants - special relations between signal parameters that remain unchanged despite changes in the signals themselves by the channel [1].

An exception is the distortion of signals by white noise, which can only be absolutely invariant by applying signals with infinitely high energy.

It can be shown that the «classical» amplitude, frequency, and relative phase modulations also use invariants of the simplest orthogonal transformation group, which preserves the length of the signal vectors and the angle difference between them.

To date, the basic properties of the invariant amplitude modulation, which uses the basic invariant of the affine transformation group, describing the entire class of linear communication channels with arbitrary frequency characteristics, have been partially studied.

At the same time, such an invariant, called in mathematics "the ratio of three points" [2], in relation to communication problems, has so far been formulated as a channel preserving the ratio of the lengths of signal vectors lying on a single straight line passing through the origin of the coordinate system of the signal space. However, the " ratio of three points "(the ratio of the lengths of segments lying on the same line and defined by these points preserved by an affine transformation) is also valid for the General case when the line does not pass through the origin [2]. The synthesis of modified invariant amplitude modulation (MIAM) and demodulation algorithms for this General case is described below.

2 SYNTHESIS OF MODIFIED INVARIANT AMPLITUDE MODULATION AND DEMODULATION

Below, in order to provide greater clarity of the synthesis procedure, a two-dimensional signal space is used, the coordinate axes of which correspond to some orthonormal basis functions of time. For example, the Kotelnikov functions $\varphi_1(t)$ and $\varphi_2(t)$, which differ in the time shift that ensures their orthogonality. In this case, the signals can be represented by two time counts, the values of which set the coordinates of the ends of the signal vectors in a two-dimensional signal space (Figure 1).

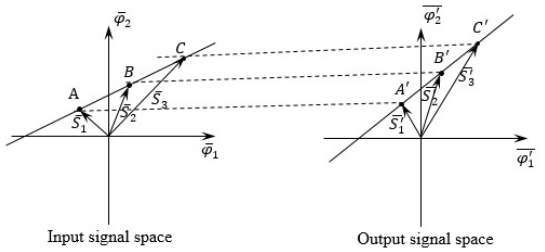


Figure 1: The scheme is an affine transformation of the input signals at the output signals for the modified invariant amplitude modulation.

Figure 1 on the left shows a straight line that occupies a general position and does not pass through the origin of the coordinate systems of the signal space. Points A, B, and C define the ends of the three input vectors $\overline{S_1}$, $\overline{S_2}$ and $\overline{S_3}$.

As mentioned above, the transformation of input signals into output signals is described by an affine transformation group. In the Figure 1 the dotted lines represent the scheme of affine transformation of the ends of the input signal vectors into the ends of the corresponding output signal vectors $\overline{S}'_1$, $\overline{S}'_2$ and $\overline{S}'_3$. One of the possible forms of writing the affine transformation invariant in the form of the "three-point relation" in this example has the following form (1) and (2).

$$J = \frac{BC}{AB} = \frac{B'C'}{A'B'} = \frac{\left|\overline{S}_3 - \overline{S}_2\right|}{\left|\overline{S}_2 - \overline{S}_1\right|} = \frac{\left|\overline{S}'_3 - \overline{S}'_2\right|}{\left|\overline{S}'_2 - \overline{S}'_1\right|}. \quad (1)$$

Such an invariant entry is also possible

$$J = \frac{AC}{AB} = \frac{A'C'}{A'B'} = \frac{\left|\overline{S}_3 - \overline{S}_1\right|}{\left|\overline{S}_2 - \overline{S}_1\right|} = \frac{\left|\overline{S}'_3 - \overline{S}'_1\right|}{\left|\overline{S}'_2 - \overline{S}'_1\right|}. \quad (2)$$

We assume that in (1) $S_1(t)$ and $S_2(t)$ perform the role of so-called "reference signals", and $S_3(t)$ – informational $S_i(t)$. In this case, the modified invariant amplitude modulation (MIAM) algorithm can be obtained from (1) - the algorithm modulation A (3):

$$\overline{S}_i = J_i\left(\overline{S}_2 - \overline{S}_1\right) + \overline{S}_2, \quad (3)$$

where:

J_i - value of the transmitted information element;

$\overline{S}_i$ - the vector of the information signal $S_i(t)$, which together with the reference signals $S_1(t)$ and $S_2(t)$ transmits the value of the information element J_i;

i - number of the time interval during which the value of the information element J_i is transmitted.

From (1) and Figure 1 follows the algorithm demodulation A (4):

$$\hat{J}_i = \frac{\left|\hat{\overline{S}}'_i - \hat{\overline{S}}'_2\right|}{\left|\hat{\overline{S}}'_2 - \hat{\overline{S}}'_1\right|} \text{ -.} \quad (4)$$

Here, the ^ sign indicates the estimates of the value J_i at the output of the demodulator and the vectors $\hat{\overline{S}}'_1$, $\hat{\overline{S}}'_2$ and $\hat{\overline{S}}'_3$ at the input of the demodulator.

Expression (2) gives other equivalent modulation and demodulation algorithms - the algorithm modulation B (5) and the algorithm demodulation B (6):

$$\overline{S}_i = J_i(\overline{S}_2 - \overline{S}_1) + \overline{S}_1, \quad (5)$$

$$\hat{J}_i = \frac{\left|\hat{\overline{S}}'_i - \hat{\overline{S}}'_1\right|}{\left|\hat{\overline{S}}'_2 - \hat{\overline{S}}'_1\right|}. \quad (6)$$

3 ESTIMATION OF NOISE IMMUNITY OF MODIFIED INVARIANT AMPLITUDE MODULATION TO WHITE NOISE

Let us denote the modules of the difference vectors for brevity $|\hat{\overline{S}}'_i - \hat{\overline{S}}'_1|$, $|\hat{\overline{S}}'_i - \hat{\overline{S}}'_2|$ and $|\hat{\overline{S}}'_2 - \hat{\overline{S}}'_1|$ as $\left|\Delta\hat{\overline{S}}'_{i,1}\right|$, $\left|\Delta\hat{\overline{S}}'_{i,2}\right|$ и $\left|\Delta\hat{\overline{S}}'_{2,1}\right|$, respectively. Then the demodulation algorithms A and B can be written as follows (7) and (8):

$$\hat{J}_i = \frac{|\Delta\hat{\bar{S}}'_{i,2}|}{|\Delta\hat{\bar{S}}'_{2,1}|}; \quad (7)$$

$$\hat{J}_i = \frac{|\Delta\hat{\bar{S}}'_{i,1}|}{|\Delta\hat{\bar{S}}'_{2,1}|}. \quad (8)$$

Taking into account the influence of white noise $n(t)$ on the transmitted signals $S_i(t)$, $S_2(t)$, and $S_1(t)$ for estimating the lengths of difference vectors $\Delta\bar{S}'_{i,1}$, $\Delta\bar{S}'_{i,2}$ and $\Delta\bar{S}'_{2,1}$ can be written (9), (10), (11)

$$|\Delta\hat{\bar{S}}'_{i,1}| = |\Delta\bar{S}'_{i,1}| + n_i - n_1; \quad (9)$$

$$|\Delta\hat{\bar{S}}'_{i,2}| = |\Delta\bar{S}'_{i,2}| + n_i - n_2; \quad (10)$$

$$|\Delta\hat{\bar{S}}'_{2,1}| = |\Delta\bar{S}'_{2,1}| + n_2 - n_1, \quad (11)$$

where n_i, n_2, n_1 - are the values of the projection lengths of the vectors of realizations of white noise interference on the direction set by the line passing through the points A', B', C' (Figure 1), and affecting the transmitted signals $Si(t)$, $S_2(t)$, and $S_1(t)$, respectively.

It is known that the projections of the white noise realization vector in any orthonormal basis are Gaussian random variables with $\sigma^2_{w.n.}$ and zero expectation [3].

With this in mind (9), (10), (11) they are Gaussian random variables with mathematical expectations $|\Delta\bar{S}'_{i,1}|, |\Delta\bar{S}'_{i,2}|$ и $|\Delta\bar{S}'_{2,1}|$ and dispersions $2\sigma^2_{w.n.}$.

Values $\hat{J}_i$ in algorithms (7) and (8) are functionally transformed random variables. As is known [4], for a functionally transformed random variable in the form $y = \frac{x_2}{x_1}$, the following inequality holds (12):

$$\omega(y) = \int_{-\infty}^{\infty} \omega(x_1, x_2)|x_1|dx_1 = \int_{-\infty}^{\infty} \omega(x_1, yx_1)|x_1|dx_1, \quad (12)$$

where $\omega(x_1, x_2 = yx_1)$ - two-dimensional law of probability distribution x_1 and x_2.

In our case, taking into account the independence of random variables in the numerators and denominators of formulas (7) and (8), we have:

$$\omega\left(|\Delta\hat{\bar{S}}'_{i,2}|, |\Delta\hat{\bar{S}}'_{2,1}|\right) = \omega\left(|\Delta\hat{\bar{S}}'_{i,2}|\right)\omega\left(|\Delta\hat{\bar{S}}'_{2,1}|\right)$$

$$\omega\left(|\Delta\hat{\bar{S}}'_{i,1}|, |\Delta\hat{\bar{S}}'_{2,1}|\right) = \omega\left(|\Delta\hat{\bar{S}}'_{i,1}|\right)\omega\left(|\Delta\hat{\bar{S}}'_{2,1}|\right)$$

In accordance with (12), we obtain the following expression for the conditional probability density of estimates of the values of information elements $\hat{J}_i$ at the output of the demodulator for algorithm A (13):

$$\omega\left(\hat{J}_i / J_i\right) = \frac{1}{4\pi\sigma^2_{w.n.}} \int_{-\infty}^{\infty} e^{-a} |\Delta\hat{\bar{S}}'_{2,1}| d |\Delta\hat{\bar{S}}'_{2,1}|, \quad (13)$$

where $a = \frac{\left(|\Delta\hat{\bar{S}}'_{2,1}| - |\Delta\bar{S}'_{2,1}|\right)^2 + \left[\left(\hat{J}_i - J_i\right)|\Delta\bar{S}'_{2,1}|\right]^2}{4\sigma^2_{w.n.}}$

For example, graphs were calculated in the MATLAB environment $\omega(\hat{J}_i / J_i)$ (Figures 2 - 5) when using reference signals with the length of the difference vector $|\Delta\bar{S}'_{2,1}|$=1 and $|\Delta\bar{S}'_{2,1}|$=2 and values of transmitted information elements J_i = 1, 2, 3, 4, 5, 6 for interference power $\sigma^2_{w.n.}$=0,1 and $\sigma^2_{w.n.}$=0,2 [5].

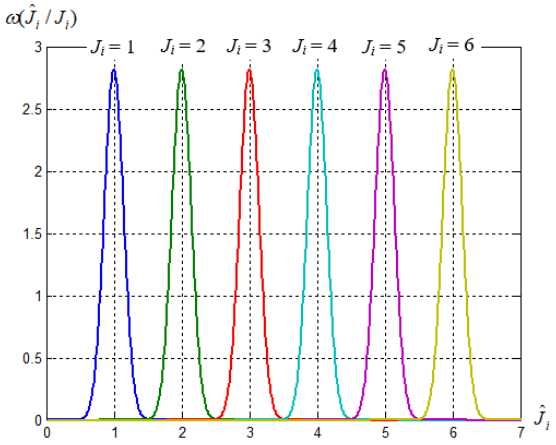


Figure 2: Graphics $\omega(\hat{J}_i / J_i)$ for $|\Delta\bar{S}'_{2,1}|$=1, $\sigma^2_{w.n.}$=0,1.

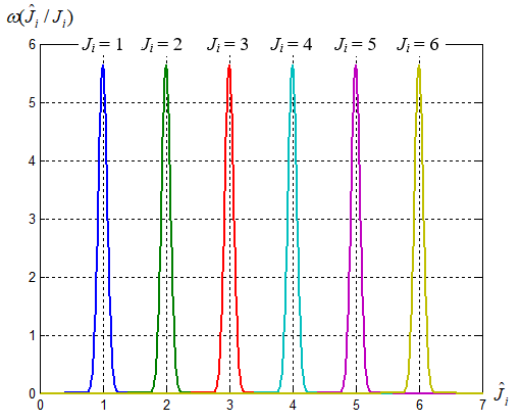


Figure 3: Graphics $\omega(\hat{J}_i / J_i)$ for $|\Delta\bar{S}'_{2,1}|$=2, $\sigma^2_{w.n.}$=0,1.

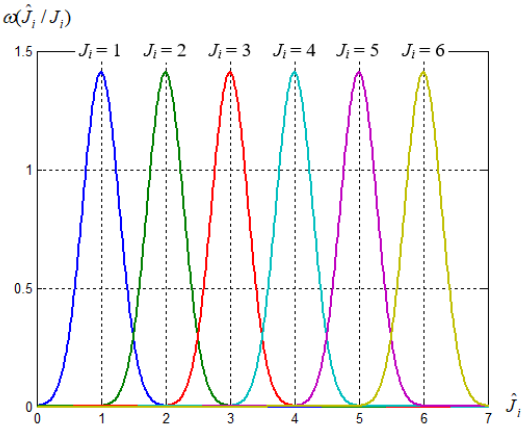


Figure 4: Graphics $\omega(\hat{J}_i / J_i)$ for $|\Delta\bar{S}'_{2,1}|$=1, $\sigma^2_{w.n.}$=0,2.

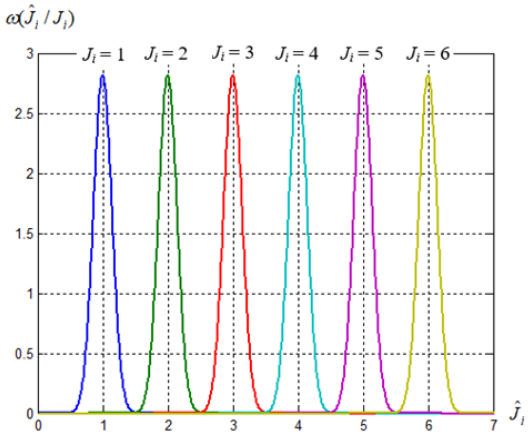


Figure 5: Graphics $\omega(\hat{J}_i / J_i)$ for $|\Delta\bar{S}'_{2,1}|$=2, $\sigma^2_{w.n.}$=0,2.

As follows from these graphs, an increase in the length of the vector of the difference between the vectors of two reference signals leads to a decrease in the variance of estimates of the values of information elements at the output of the demodulator. Therefore, by choosing the required length of the difference vector, it can be provided the necessary minimization of the area of mutual overlap of graphs $\omega(\hat{J}_i / J_i)$, that is, the required fidelity of the J_i transmission.

3 CONCLUSIONS

The proposed communication system with modified invariant modulation is a further generalization of the known system with invariant amplitude modulation [1].

The advantage of this communication system is that there is no need to ensure the similarity of the forms of the signals used, as is the case with the prototype. The line of location of the ends of the vectors of transmitted signals can occupy an arbitrary position. This circumstance can be used to increase the degree of secrecy of transmitted messages, similar to the well-known method of carrier frequency tuning.

In addition, when changing the sequence of transmission of reference signals, taking into account the lengths of their vectors, it becomes possible to transmit a sign of positivity or negativity of the values of information elements. For example, a positive sign may be that the first time reference signal has a shorter vector length compared to the vector length of the second time reference signal. The negative sign is transmitted in the reverse order of the reference signals [5].

Thus, the proposed modified invariant modulation allows to double the volume of the alphabet of information elements compared to the prototype.

REFERENCES

[1] V. V. Lebedyancev, “Development and research of methods for analysis and synthesis of invariant communication systems,” Dissertation for the degree of doctor of technical sciences, Novosibirsk, 1995.

[2] N. V. Efimov, Higher geometry, Moscow, Nauka, 1978, 576 p.

[3] B. R. Levin, Theoretical foundations of static radio engineering, 1st ed., Moscow, Sovetskoe radio, 1969, 752 p.

[4] A. M. Zajezdnii, Fundamentals of calculations for static radio engineering, Moscow, Communication, 1969, 448 p.

[5] V. V. Lebedyancev, S. S. Abramov, I. I. Pavlov, E. V. Morozov, E. S. Abramova, and M. S. Pavlova, "Modified invariant amplitude modulation," Journal T-Comm. Telecommunications and transport. 2020, t.14, no. 6, pp. 13-19.

Method for Determining the Number of States of the Markov Model of Damage Accumulation in Predicting the Technical Condition of a Fiber-Optic Cable

Elena Ionikova[1], Kirill Karpov[1,2] and Viacheslav Shuvalov[1]

[1] *Department of Transmission of Discrete Data and Metrology, Siberian State University of Telecommunications and Information Sciences, Kirova Str. 86, 630102 Novosibirsk, Russian Federation*

[2] *Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*

ionikova.lena@mail.ru, kirill.karpov@hs-anhalt.de

Keywords: Service Life, Markov Chain, Gradual Failures, Weibull Parameters, Approximation.

Abstract: Estimation of the residual service life of operating fiber-optic cables is an urgent t ask. Usually this problem is solved based on the use of the Markov chain model. However, due to the nonlinear dependence of the probability of rejection on the rate of gradual failures, the task of selecting the number of states of a Markov chain becomes difficult. The article discusses a technique for determining the required number of states of the Markov model of damage accumulation based on a given value of the modeling accuracy. The characteristic values of the time and the probability of failure are found for the model of the destruction of optical fibers made of silica glass. The determination of the required number of states of a Markov chain in the article is carried out using the Python programming language.

1 INTRODUCTION

Since 1993, more than 2.2 billion km of optical fiber has been laid in the world, which is used to transmit more than 20% of global information traffic. With the proliferation of cloud technologies, distributed computing and databases, the role of fiber-optic communication systems is growing steadily [1].

The service life of fiber-optic communication lines is about a quarter of a century [2, 3] (not to be confused with the warranty service life, which, according to most manufacturers, does not exceed two years). Depending on the design of the cable and its field of application, the value of the service life can vary from 2 to 40 years. Often the requirements for tendering indicate that the service life of an optical cable must be at least 25 years.

Lifetime of the cable is defined [4, 5] as the average service life - the mathematical expectation of the service life. Gamma percent service life is the calendar duration of operation during which it will not reach the limit state with a given probability γ, expressed as a percentage.

A. Yu. Tsym in his work [6] proposed to supplement the list of lifetime criteria with an indicator of disproportionate risk of loss of network connectivity.

This indicator is relevant for the Russian information infrastructure due to the limited possibilities of network redundancy and the need for an additional assessment of the fact of loss of network connectivity when the optical cable goes to the limit state. The criteria for the limiting state is a set of features established in the technical documentation [7].

During its service life, a fiber-optic cable belongs to the class of recoverable objects, after passing to the limit state, it is a non-recoverable object. The transition to the limiting state occurs gradually as the static fatigue of the fiber accumulates (aging or deterioration).

In addition to damage to the cable sheath, the aging of optical fibers is influenced by such internal factors as fiber stretching, moisture, and hydrogen [8, 9]. The lifetime of optical cables is mainly determined by the amount of tension on the fibers. Since under tension, optical fibers gradually decrease their strength due to the growth of cracks on their surface [10, 11], the number of failures caused by cable breaks increases. For example, the ITU-T recommendation [11] provides test data for cable sections that have been buried in the ground since 1979, 1986 and 1991. The probability of failure values for the 1979 cable are significantly higher than those for the 1991 cable. In the work of I. V. Bogachkov and N.I. Gorlov [12] it is shown that the established service life of 25 years is

ensured in the presence of an elongation of less than 0.26%, which determines the permissible value of the local mechanical tensile load within 3 N.

2 RELATED WORK

Gradual failure models are designed to analyze changes in the physical parameters of technical systems under stress. The parameter Y, called *defining*, changes during wear, reaching a limit value, after which the system becomes inoperative. The mean time to failure of the system is determined by the formula (1).

$$T = \frac{Y_{lim} - Y_0}{\alpha} \tag{1}$$

where Y_{lim} — limiting value of the defining parameter; Y_0 — its initial value; α — the rate of change of the defining parameter $\frac{dY}{dt}$.

Many works have been devoted to assessing the lifetime of various objects, including fiber-optic lines, in which various models of the transition of an object to the limiting state are described.

It was shown in [13] that the processes of damage accumulation (regardless of their nature) can be described by Markov models, on the basis of which it is possible to construct fairly accurate models of cumulative damage accumulation. In [14], this model was used to describe the accumulation of damage in polymer high-voltage insulation. In [15], this probabilistic approach was used to model the life cycle of road bridge elements based on the Markov stochastic degradation model. The paper presents a graph of the degradation process for a model of five discrete states and a method for determining the degradation parameter, which is considered as the failure rate λ.

In [16], the Markov model with discrete states and continuous time is used to predict the parametric reliability of the Monitoring System. Application of this model makes it possible to determine several operational states of the Monitoring System with different levels of operational efficiency, determined by the probability of no-failure operation.

In [17], a Markov branching process was used to build a model for predicting changes in the parameters of an electronic system during operation. The model is recommended to be used to predict the parametric reliability and technical condition of radio-electronic systems depending on the time of operation.

In [18], Markov models were developed for predicting the parameters of computer networks, taking into account the nonstationarity of the operation modes. The calculation of the parameters is carried out on the basis of the results of the wavelet - analysis of the dynamics of changes in operating parameters.

In [19], using the theory of semi-Markov processes, models of operation of communication systems equipment are considered, taking into account the physical aging of the elements included in it.

A similar approach to assessing the time to reach the limit state can be used for fiber-optic communication lines.

Griffiths is considered the founder of the mechanical concept of optical fiber destruction [20, 21]. According to Griffiths, a solid contains microcracks, which begin to expand under the action of tensile stress. Crack growth occurs when the tensile force reaches a certain threshold value. When this value is reached, the crack begins to grow at a limiting rate.

Today, optical fiber fracture models are actively used, built on the basis of the empirical concept of the power-law dependence of the rate of development of microcracks V on the tensile stress intensity factor K_e, which characterizes the overstress at the crack tip.

$$V = A \cdot K_e^n \tag{2}$$

where n is the parameter of resistance to fatigue (corrosion coefficient); A is a constant depending on the parameters of the material and the environment.

In [22], it is noted that the use of a simple power law to describe the statistical fatigue of an optical fiber leads to the neglect of the possible existence of regions where crack growth follows other mechanisms and patterns (regions with a limited rate of moisture diffusion to the crack tip, as well as regions of thermal fluctuation crack growth in the absence of moisture).

A similar approach to estimating the time to reach the limit state can be used for fiber-optic communication lines.

This article discusses a technique for determining the required number of states of a Markov damage accumulation model based on a given value of modeling accuracy. The characteristic values of the time and the probability of failure are found for the model of the destruction of optical fibers made of quartz glass according to the Weibull parameters determined in the article [23].

3 GRADUAL FAILURE MODELS

The theoretical aspects of mechanical reliability are described in sufficient detail in the document [24] and articles [25, 26], where the following formula for calculating the time to failure for static fatigue is presented:

$$t = \frac{2}{A \cdot Y^2 (n-2) \cdot K_{IC}^{n-2}} \cdot \frac{\sigma_c^{n-2}}{\sigma_{exp}^n} \tag{3}$$

Where σ_c — fiber strength in an inert environment; σ_{exp} — applied tension; A — a constant depending on the material and the environment; Y — coefficient depending on the geometry of the crack; n — fatigue parameter; K_{IC} — the stress intensity factor corresponding to the inert environment.

For a statistical assessment of the mechanical strength of an optical fiber, the most suitable type of distribution is the Weibull law, written in the form [27]:

$$P(\sigma, L) = 1 - \exp\left(-\frac{L}{L_0} \left(\frac{\sigma}{\sigma_0} \right)^n \right) \tag{4}$$

where L is the length of the optical fiber; L_0 — the length of the optical fiber sample during testing; σ — tensile strength of the fiber; σ_0, m — the parameters of the Weibull distribution are determined experimentally.

At present, usually, a two-stage optical fiber destruction model is used: the first mode is valid for the probabilities of optical fiber destruction $P_{crit} \leq P(\sigma, L) \leq 1$, the second is for probabilities $0 \leq P(\sigma, L) \leq P_{crit}$. P_{crit} corresponds to the probability of destruction at the boundary of two modes.

$$P(\sigma, L) = \begin{cases} 1 - \exp\left(\frac{L\sigma^{m_1}}{L_0 \sigma_{01}^{m_1}} \right) & if\ P_{crit} \leq P(\sigma, L) \\ 1 - \exp\left(\frac{L\sigma^{m_2}}{L_0 \sigma_{02}^{m_2}} \right) & if\ P(\sigma, L) \leq P_{crit} \end{cases} \tag{5}$$

However, for gradual failures, it is more reasonable to consider only the time interval of the two-stage fiber failure model, characterized by a slow decrease in the availability factor (a slow increase in the probability of failure).There is no need to take into account the time interval corresponding to the transition to the limit state, since the second stage proceeds in fractions of a second [28] and the value of the unavailability coefficient tends to 1.

From equations (3) and (4) it follows that the probability of failure of an optical fiber during its aging is determined as:

$$P(t) = 1 - e^{-\frac{L}{L_0} \cdot \left[\frac{\sigma_c}{\sigma_0} - \left(\left[\frac{\sigma_c^{n-2}}{\sigma_0^{n-2}} \right] - \frac{t \cdot \sigma_{exp}^n}{B \cdot \sigma_0^{n-2}} \right)^{\frac{1}{n-2}} \right]^m} \tag{6}$$

where

$$B = \frac{2}{A \cdot Y^2 (n-2) \cdot K_{IC}^{n-2}} \tag{7}$$

The Markov model of damage accumulation with discrete states and continuous time can be represented as Figure 1.

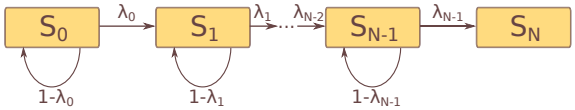


Figure 1: Markov Damage Accumulation Model.

The probability of the transition of the system to the state S_j during the time interval Δt, counted from the moment t, will be denoted by $P_{ij}(t + \Delta t)$.

$$P_{ij}(t + \Delta t) = P(S(t + \Delta t) = S_j | S(t) = S_i) \tag{8}$$

In this case, the events of the Markov chain are mutually exclusive and create a complete group:

$$\sum_{k=1}^{N} P_k(t) = 1 \tag{9}$$

The probability density of the transition (or the intensity of the transition) of the system from state S_i to state S_j is:

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{P_{ij}(t + \Delta t) - P_{ij}(t)}{\Delta t} = P'_{ij}(t) \tag{10}$$

Note that for $\Delta t \to 0$ the intensity $\lambda_{ij}(t) = \tan(\alpha)$, where $\tan(\alpha)$ is the tangent to the function $P_{ij}(t)$.

$$P_{ij}(t + \Delta t) \approx \lambda_{ij} \Delta t \tag{11}$$

Thus, the intensities of state-to-state transitions can be found by means of a piecewise linear approximation of the failure probability function determined by formula (6). The accuracy of the correspondence of the piecewise linear approximation of the original function will depend on the number of linear sections of the approximating function, the number of which will correspond to the number of states of the Markov model. The choice of the optimal number of states in this case is determined by the required simulation accuracy.

4 DATA FOR MODELING

The following characteristics are taken as the initial data for finding the required number of states of the Markov damage accumulation model for an optical fiber 100 km long:

1) $L_0 = 0.012$ m, $\sigma_0 = 5.222$ GPa, $m = 5.187$, $n = 23.287$, $\ln B = -24.7711$ [23]

2) Inert strength for category B singlemode fibers $\sigma_c = 20$ APa according to [29];

3) Applied Aechanical Atress $_{exp}$=2 APa Taking into account the adopted characteristics, expression (6) takes the form:

$$P(t) = 1 - e^{-\left[53.0476-(5.164\cdot 10^{36}-6.1\cdot 10^{26}\cdot t)^{\frac{1}{21.287}}\right]^{5.187}} \quad (12)$$

5 FINDING THE REQUIRED NUMBER OF STATES OF THE MARKOV MODEL

Piecewise-linear approximation of the function of the probability of failure of an optical fiber at any moment of time is obtained in the following form:

$$P(t) = \begin{cases} k_1 t + C_1 & if\, 0 \le t \le t_1 \\ k_2 t + C_2 & if\, t_1 \le t \le t_2 \\ \dots & \dots \\ k_v t + C_v & if\, t_{v-1} \le t \le t_v \end{cases} \quad (13)$$

where v is the number of sections in the piecewise linear approximating function, the number of states of the Markov model of damage accumulation.

To find the number of states of the Markov model of damage accumulation, it is necessary to set the permissible simulation error ε_{mod}, which will correspond to the approximation error ε_{approx}.

$$\varepsilon_{mod} \ge \varepsilon_{approx} \quad (14)$$

The solution of the problem of approximation with the required error within the framework of the article is implemented by numerical methods.

Finding the optimal linear equations for piecewise linear approximation for a function is carried out using the least squares method [30]. The software functionality is implemented in the pwlf library for the Python programming language. The approximating function is found for a given number of linear equations. To determine the required number of linear equations, the criterion of the maximum approximation error ε_{approx} is used for a time interval from N_1 to N_2 days. The number of approximating linear equations increases until condition (14) is satisfied. The maximum error between the approximating and approximating functions in percent is found as:

$$dmax(f, \pi f) = \max(f - \pi f) \cdot 100 \quad (15)$$

where f is the approximated function, $\pi\, f$ - approximating function.

As an example, an algorithm of operation is presented at ε_{mod} = 0.005, for a time interval from 1 year to 60 years.

```
import pwlf
import numpy as np

P(t) # equation 11

e_max_threshold = 0.005
nLine = 1
t0, t1 = 3.1536e7, 1.89216e9
t = np.linspace(t0, t2, 10000)

maxError = lambda (f, pf) :
           max(abs(f - pf)) * 100

pf = pwlf.PiecewiseLinFit(t, P(t))

while (e_max > e_max_threshold):
    nLine += 1
    pf.fit(nLine)
    e_max = maxError(P(t),
                pf.predict(t))
```

Based on the results of the algorithm, an approximating function is determined, consisting of v linear equations with the required approximation accuracy.

6 RESULTS

The article discusses the dependence of the number of states of the Markov model of damage accumulation on the required modeling error. Considered modeling errors: 0.0001, 0.0005, 0.001, 0.005, 0.01 and 0.05. The obtained values are presented in Table 1. The dependence of the approximation accuracy ε_{approx} on the number of states of the Markov model is shown in Figure 6.

The plots of the approximating functions *vs* the absolute approximation error, are shown in Figures 2-5.

Table 1: Approximation results with different modeling errors.

ε_{mod}	ε_{approx}	# of states	Equation
0.0005	0.00048	9	(16)
0.001	0.00095	6	(17)
0.005	0.0038	3	(18)
0.01	0.009	2	(19)
0.05	0.009	2	(19)

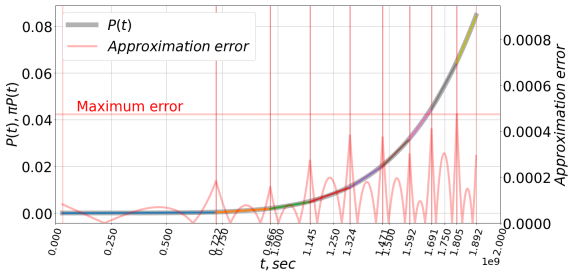


Figure 2: The plot of the approximating function combined with the absolute error of approximation with the accuracy of modeling $\varepsilon_{mod} = 0.0005$.

$$\pi P(t) = \begin{cases} 4.3 \cdot 10^{-13} x & if\ 3.2e7 > t \geq 7.2e8 \\ 6.5 \cdot 10^{-12} x - 0.004 & if\ 7.2e8 > t \geq 9.7e8 \\ 1.6 \cdot 10^{-11} x - 0.014 & if\ 9.7e8 > t \geq 1.1e9 \\ 3.4 \cdot 10^{-11} x - 0.034 & if\ 1.1e9 > t \geq 1.3e9 \\ 6.2 \cdot 10^{-11} x - 0.07 & if\ 1.3e9 > t \geq 1.5e9 \\ 9.5 \cdot 10^{-11} x - 0.12 & if\ 1.5e9 > t \geq 1.6e9 \\ 1.3 \cdot 10^{-10} x - 0.17 & if\ 1.6e9 > t \geq 1.7e9 \\ 1.7 \cdot 10^{-10} x - 0.25 & if\ 1.7e9 > t \geq 1.8e9 \\ 2.2 \cdot 10^{-10} x - 0.34 & if\ 1.8e9 > t \geq 1.9e9 \end{cases} \quad (16)$$

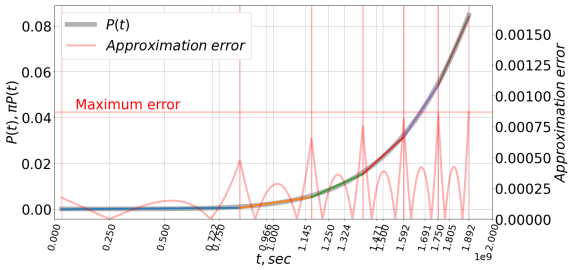


Figure 3: The plot of the approximating function combined with the absolute error of approximation with the accuracy of modeling $\varepsilon_{mod} = 0.001$.

$$\pi P(t) = \begin{cases} 8.0 \cdot 10^{-13} x & if\ 3.2e7 > t \geq 8.5e8 \\ 1.4 \cdot 10^{-11} x - 0.01 & if\ 8.5e8 > t \geq 1.2e9 \\ 4.3 \cdot 10^{-11} x - 0.04 & if\ 1.2e9 > t \geq 1.4e9 \\ 8.6 \cdot 10^{-11} x - 0.1 & if\ 1.4e9 > t \geq 1.6e9 \\ 1.4 \cdot 10^{-10} x - 0.2 & if\ 1.6e9 > t \geq 1.8e9 \\ 2.1 \cdot 10^{-10} x - 0.3 & if\ 1.8e9 > t \geq 1.9e9 \end{cases} \quad (17)$$

$$\pi P(t) = \begin{cases} 3.2 \cdot 10^{-12} x - 0.001 & if\ 3.2e7 > t \geq 1.2e9 \\ 5.8 \cdot 10^{-11} x - 0.06 & if\ 1.2e9 > t \geq 1.6e9 \\ 1.7 \cdot 10^{-10} x - 0.2 & if\ 1.6e9 > t \geq 1.9e9 \end{cases} \quad (18)$$

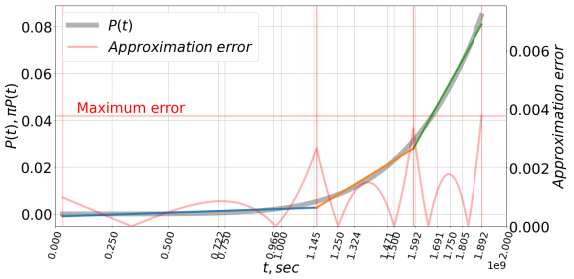


Figure 4: The plot of the approximating function combined with the absolute error of approximation with the accuracy of modeling $\varepsilon_{mod} = 0.005$.

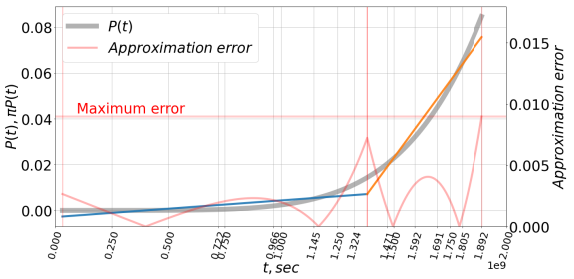


Figure 5: The plot of the approximating function combined with the absolute error of approximation with the accuracy of modeling $\varepsilon_{mod} = 0.01$.

$$\pi P(t) = \begin{cases} 7.2 \cdot 10^{-12} x - 0.003 & if\ 3.2e7 > t \geq 1.4e9 \\ 1.3 \cdot 10^{-10} x - 0.17 & if\ 1.4e9 > t \geq 1.9e9 \end{cases} \quad (19)$$

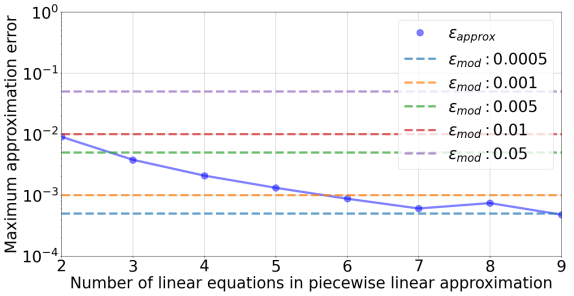


Figure 6: The plot of the dependence of the number of states of the Markov model on the required modeling error ε_{mod}.

Thus, using this algorithm, it is possible to determine the number of states of the Markov damage accumulation model for a given value of the simulation accuracy. In this case, the intensities of transitions between the states of the Markov model can be found from the obtained system of equations of approximating functions.

For a modeling error of 0.01, only two equations are enough to describe a given function, for a model-

ing error of 0.005, three equations are enough. However, such accuracy is often unacceptable with the required reliability indices of 0.982 for the backbone primary network, 0.998 for the intra-zone primary network [31]. At the same time, with an increase in the number of equations (when there are more than 8), the modeling error decreases slightly. The optimal value in this case is modeling with an error of 0.001.

However, the accuracy of the modeling should be determined by the infrastructure owner based on many aspects: the category of users, the economic costs of downtime, etc.

7 CONCLUSIONS

The article proposes a method for solving the problem of determining the number of states of a Markov chain associated with a nonlinear dependence of the probability of rejection on the rate of gradual failures. From the system of equations of approximating functions, the values of time and probability of failure can be found for the model of failure of optical fibers made of silica glass.

REFERENCES

[1] I. Dezhina, A. Ponomarev, Gabitov, and team of authors, “Public Analytical Report on Development of Photonics in Russia and the World,” Moscow, Tech. Rep., 2016.

[2] “Frequently Asked Questions on Fiber Reliability,” Corning, White Paper, April 2016.

[3] R. Grunvalds, A. Ciekurs, J. Porins, and A. Supe, “Evaluation of Fibre Lifetime in Optical Ground Wire Transmission Lines,” Latvian Journal of Physics and Technical Sciences, publisher: De Gruyter Poland, vol. 54, no. 3, p. 40, 2017.

[4] “GOST 27.002-2015 Dependability in technics. Terms and definitions,” INMIT, GOST, 2016.

[5] V. Ostreykovsky, “Reliability theory. textbook.” 2012.

[6] A. Y. Tsym, “Service life of optical cables. analyzes. risks.” IV Russian Scientific and Technical Conference ”Communication Lines of the XXI Century”, 2019.

[7] M. A. Karapuzov, S. N. Polessky, I. A. Ivanov, and P. S. Korolev, “Evaluation of livetime indicators of radioelectronic devices,” T-Comm, vol. 9, no. 7, 2015.

[8] A. Listvin and V. Listvin, “Reflectometry of optical fibers,” 2005.

[9] ITU, “Series g: Transmission systems and media, dig-ital systems and networks guidance on optical fibre and cable reliability-itu-t g-series recommendations –supplement 59,” Tech. Rep.

[10] V. Gurtov, “Optoelectronics and fiber optics: A study guide,” Ed. PetrSU, 2005.

[11] G. S. Glaesemann, “Optical fiber mechanical reliability,” vol. 8002, pp. 1-62.

[12] I. V. Bogachkov, A. I. Trukhina, and N. I. Gorlov, “Detection of optical fiber segments with mechanical stress in optical cables using brillouin reflectometers,” in 2019 International Siberian Conference on Control and Communications (SIBCON), pp. 1-7, ISSN: 2380-6516.

[13] F. Kozin and J. L. Bogdanoff, “Probabilistic models of fatigue crack growth: results and speculations,” Nuclear engineering and design, publisher: Elsevier, vol. 115, no. 1, pp. 143-171, 1989.

[14] Z. Uzagaliev, “Probabilistic model of damage accumulation in the polymer high voltage insulation,” Vesnik KRSU, vol. 15, no. 9, pp. 156-159, 2015.

[15] A. I. Lantuh-Lyaschenko, “On the issue of ”memory” markov model of damage accumulation,” Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport, no. 33, 2010.

[16] S. Fadin, K. Krasyuk, and R. Trishch, “Markov prediction model of parametric reliability of measuring and computing systems,” Eastern-European Journal of Enterprise Technologies, vol. 2, no. 6 (38), 2009.

[17] A. Grishko, “Stochastic model of parametric prediction of reliability of radio-electronic systems,” Information-measuring and control systems, no. 2(28), pp. 5-12, 2019.

[18] L. Tereikovska, I. Tereikovskyi, I. Parkhomenko, and Toliupa, “Markov model of normal conduct template of computer systems network objects,” 2018.

[19] E. Mironov, I. Mishchenko, and S. Platonov, “Creation of communication systems equipment operation models considering aging,” T-Comm, vol. 12, no. 6, 2018.

[20] A. A. Griffith and J. J. Gilman, “The phenomena of rupture and flow in solids,” Transactions of the ASM, vol. 61, pp. 855-906, 1968.

[21] R. Sack, “Extension of griffith’s theory of rupture to three dimensions,” Proceedings of the Physical Society (1926-1948), vol. 58, no. 6, p. 729, 1946.

[22] S. Semenov, “Reliability, durability, aging and degradation of silica glass fibers.abstract of the dissertation for the degree of doctor of physical and mathematical sciences,” 2007.

[23] A. V. Milkov and M. Yakovlev, “Reliability assessment of optical fibers based on short-term strength and static fatigue testing,” Systems and means of communication, television and radio broadcasting, no. 1-2, 86, 2002.

[24] T. Volotinen, M. Gadonna, H. Limberger, and team of authors, “Reliability of optical fibers and components: achievements and conclusions of COST 246,” in Optical Fiber Reliability and Testing, vol. 3848. International Society for Optics and Photonics, 1999, pp. 88-94.

[25] M. J. Matthewson, “Strength-probability-time diagrams using power law and exponential kinetics models for fatigue,” in Reliability of Optical Fiber Components, Devices, Systems, and Networks III, vol. 6193. International Society for Optics and Photonics, 2006, p. 619301.

[26] L. K. Baker, “Comparison of Mechanical Reliability Models for Optical Fibers,” Corning White Paper WP5049, 2001.

[27] A. G. Evans and S. M. Wiederhorn, “Proof testing of ceramic materials - an analytical basis for failure prediction,” International Journal of fracture, publisher: Springer, vol. 10, no. 3, pp. 379-392, 1974.

[28] A. V.Andreev, V.Burdin, “Scenarios of prediction optical fiber lifetame in cable lines,” Last Mile, no. 4, pp. 34-43, 2020.

[29] “IEC 60793-1-33:2001 Optical fibres. Part 1-33. Methods of measurement and testing. Corrosion resistance in stressed state,” Tech. Rep., 2001.

[30] N. Golovchenko, “Least-squares fit of a continuous piecewise linear function,” 2004.

[31] “RD 45.047-99 Fiber optic transmission lines for backbone and intraareal primary networks of all-Russia telecommunication system. Technical operation,” Tech. Rep., 1999.

Distributed Fiber-Optic Sensors Based on Principle of Stimulated Brillouin Scattering

Igor Bogachkov[1], Nikolai Gorlov[2] and Evgenia Kitova[3]
[1]*Department of Communications and Information Security, Mira Str. 11, Omsk State Technical University, 644050 Omsk, Russian Federation*
[2]*Department of Communication Lines, Siberian State University of Telecommunications and Computer Science, Kirov Str. 86, 630102 Novosibirsk, Russian Federation*
[3]*Department of Foreign Languages, Novosibirsk State Technical University, Karl Marx avenue 20, 630073 Novosibirsk, Russian Federation*
bogachkov@mail.ru, gorlovnik@yandex.ru, kitovaet@mail.ru

Keywords: Optical Fiber, Brillouin Scattering, Distributed Temperature Sensors, Optical Fiber Sensors.

Abstract: The report is devoted to the analysis of distributed fiber-optic sensors based on the phenomenon of stimulated Brillouin scattering. They are of great interest for research due to their ability to measure the temperature and strains at superlong distances with high accuracy and high spatial resolution. The functional dependences of the output signals characteristics on parameters measured by the sensors are given in the paper. Of particular interest are the results of the analysis of the spectral component shifts in the Brillouin light scattering depending on the fiber elongation and temperature. After a brief review of the basic theoretical principles the results of some researches aimed to expand the dynamic range and to increase spatial resolution. The results of simulation in professional design software environment OptiSystem 17.1 are described in the article. To test the simulation results and detection of common features in spectrograms the experimental testing was carried out. The results obtained show that it is possible to implement fiber-optic sensors based on Brillouin scattering in telecommunication systems, mining, oil and gas industry, as well as in electric-power industry, construction, aviation and space industry. The objectives for the further research are to perform metrological analysis at all stages of the method implementation, to complete the base of Brillouin spectrograms for optical fiber of various types and to improve algorithms for automated processing spectra in order to expand functionality of the systems. In conclusion, the overview of some applications is given in this paper.

1 INTRODUCTION

Optic fibers are widely used as communication channels in which light waves can be transmitted over long distances. In this situation fiber lines are isolated from external disturbances by means of cable technologies. However, by increasing environmental influences on the properties of light penetrating into the waveguide, fiber can be used for detection, monitoring and even measurement of external disturbances (measured values) in the integrated or distributed format. When optical power exceeds the prescribed power threshold, nonlinear phenomena, such as Brillouin scattering due to its strong dependence on external environmental variables (deformation and temperature), then the waveguide can be used successfully in optic sensor systems. In these cases, the optic fiber is a medium in which interaction occurs, acting simultaneously as both a distributed converter and an optical channel. These sensors can measure changes in particular parameter along the whole fiber converter. Therefore, dynamic range, correlated to the maximum fiber length and spatial resolution of the converter (minimal fiber length required for measurement during serial disturbances or events) are key factors which are significant and they should be investigated [1].

Fiber-optic sensors have the following advantages compared to their traditional counterparts [2]:

- fiber-optic line is explosion- and fireproof;

- optic fiber has high resistance to the influence of corrosive media, pressures and temperatures;
- optical signal in the fiber sensor is not affected by electric of magnetic interference caused by the operation of other technical systems;
- fiber sensors are distributed data collection system with the remote information processing devices;
- they easily operate in combination with optical information processing systems;
- small size and weight of optic fiber;
- high corrosion resistance, especially to chemical solvents, oil and water;
- free of induction;
- low cost.

The above advantages make it possible to design, manufacture and technical operation of distributed systems for monitoring of long lines along the whole length in real-time (overhead and underground communication lines, electric power lines, bridges, dams), where it is necessary to control the strength of the structure and the risk of an emergency. In addition, these systems successfully solve the problem of measuring the temperature along the entire well drilling for a long period.

2 THEORETICAL BACKGROUND

A distributed fiber-optic sensor is a kind of measuring instrument. It transforms a physical quantity which is measured into the optical signal. This signal is transmitted through the optical fiber to a processing device to process the optical signals. Algorithms used for processing the intermediate data received are provided by the required metrological sensor characteristics. Furthermore, they provide the information presentation in tabular or graphical format.

The measurement principles used in the proposed fiber-optic sensor are based on the Mandelstam-Brillouin scattering (scattering by acoustic phonons) in the optical fiber. Brillouin scattering result in the formation of the backward wave in the fiber. By scanning the carrying frequency of this wave, you can determine the distribution of the Brillouin scattering spectrum along the fiber and, consequently, the maximum signal frequency in this spectrum, Figure 1 [3].

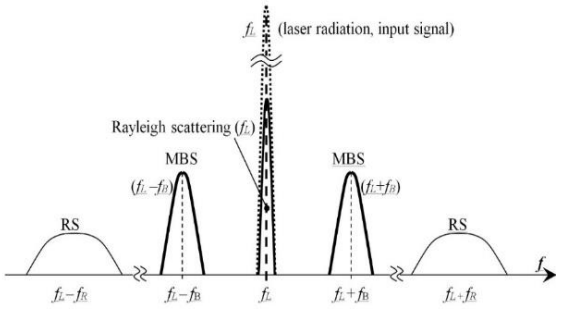


Figure 1: Spectrum of light scatterings in the fiber (fB ~ 10 ... 11 GHz, fR ~ 13 THz).

The system of the distributed sensing, which is based on the Brillouin scattering can simultaneously measure temperature and strain along the fiber. Typically, the intensity of light and the frequency shift in the Brillouin scattering is affected by the temperature and deformation in the fiber. These are defined by the following (1) [4]:

$$\begin{bmatrix}\Delta v_B(T,\varepsilon)\\ \Delta P_B(T,\varepsilon)\end{bmatrix}=\begin{bmatrix}C_{v,T} & C_{v,\varepsilon}\\ C_{P,T} & C_{P,\varepsilon}\end{bmatrix}\begin{bmatrix}\Delta T\\ \Delta\varepsilon\end{bmatrix}, \quad (1)$$

where:

$\Delta P_B\ (T, \varepsilon)$ – variations of power in the Brillouin scattering spectrum;

$\Delta \nu_B\ (T, \varepsilon)$ – shift of the central frequency in the Brillouin scattering;

ΔT – fluctuations in fiber temperature;

$\Delta\varepsilon$ – fluctuations in fiber strain;

$C_{P,T}$ and $C_{P,\varepsilon}$ are temperature coefficients and the strain coefficient for the Brillouin scattering power respectively;

$C_{V,T}$ and $C_{V,\varepsilon}$ are also the coefficients of temperature and strain for the shift of the Brillouin frequency depending on temperature and strain.

These preset values of the coefficients are given in Table 1:

Table 1: The relationship of Brillouin light scattering intensity and frequency shift and temperature coefficient.

	Strain coefficient	Temperature coefficient
Brillouin scattering optical power variation ΔP_B	$C_{P,\varepsilon} = -8 \pm 1 \times 10^{-4}$%/με	$C_{P,T} = +0.33 \pm 0.3$ % /K
Brillouin scattering optical frequency shift ΔV_B	$C_{V,\varepsilon} = -+0.052 \pm 0.004$ MHz/με	$C_{V,T} = +1.09 \pm 0.083$ MHz / K

Linear coefficient for temperature and strain can be accurately determined from the inverse matrix in (1) and can be written as [5]:

$$\Delta T = \frac{|C_{P,\varepsilon}|\cdot\Delta\nu_B + |C_{\nu,\varepsilon}|\cdot\Delta P_B}{|C_{P,T}C_{\nu,\varepsilon} - C_{P,\varepsilon}C_{\nu,T}|} \quad (2)$$

$$\Delta\varepsilon = \frac{|C_{P,T}|\cdot\Delta\nu_B + |C_{\nu,T}|\cdot\Delta P_B}{|C_{P,T}C_{\nu,\varepsilon} - C_{P,\varepsilon}C_{\nu,T}|} \quad (3)$$

By measuring the power of the scattered signal and the shift of the Brillouin frequency, it is possible to obtain the temperature and strain distribution along the fiber.

Distributed Brillouin fiber-optic sensors provide innovative solutions to control the temperature and strain in distributed constructions. The effective range for these sensors is about 20-30 km, which is limitation for their use in some applications where the distance is much longer. In order to increase the operating range, the methods are proposed in paper [6], based on distributed Raman amplification. Three Raman pumping configurations were investigated theoretically and experimentally: joint propagation, counter propagation and bidirectional propagation with respect to the Brillouin pump pulse. The study shows that some of the amplification schemes tested can significantly extend the measurement range and improve the quality of measurements over large distances.

The paper [7] proposes a combined amplification of the second order and optical pulse coding to expand the real dynamic range for distributed fiber-optic sensor. The analysis presented and the experimental results show that the appropriate optimization of these two methods makes it possible to enhance the signal/noise ratio measurements when using a superlong and sensitive fiber. This solution increases the sensing distance to 120 km with a spatial resolution of 5 m.

In order to increase the spatial resolution, the paper [8] proposes and shows a new differential method of optical reflectometry. By analyzing spatiotemporal property of the pulse excited by Brillouin spontaneous light scattering, it made it possible to obtain the distribution of the Brillouin weighting coefficient along the fiber. Based on this distribution a method of two-step subtraction is offered. A pair of pulses with a small difference in width is used as a probe pulse. When performing the two-stage subtraction on these two pairs of Brillouin spectrum it is proved theoretically and experimentally that the differential Brillouin spectrum is spatially associated with the difference in width of pulse pair. The spatial resolution of 0.4 m is obtained experimentally when using pulse pairs 60/56 ns for the probing length of 7.8 km with accuracy of the Brillouin frequency of 4.1 MHz.

The analysis results of the noise by using Monte-Carlo method to correct the dependences between the frequency resolution, quality, signal/noise ratio and the frequency step in the distributed Brillouin fiber-optic sensors are presented in [9]. Quantitative estimation of the Brillouin amplification spectrum is of great importance for the distributed sensors to increase the Brillouin frequency resolution and corresponding Brillouin tension and temperature resolutions. To estimate the error in determination of the Brillouin central frequency spectrum two analytical expressions were obtained with polynomial second order fitting and without it.

In paper [10], an equivalent Rayleigh criterion is offered to measure stressed cross section and smaller spatial resolution for the Brillouin distributed sensor. According to this criterion, at any preset probing length the minimum allowable tension length is 1/2 of the pulse length at Brillouin frequency uncertainty of 5%.

3 RESULTS AND DISCUSSION

The simulation was carried out in the professional design environment OptiSystem 17.1. It is a comprehensive program software Design Suite, which allows users to plan, test and simulate processes in optical fibers in current telecommunication systems. The computer simulation scheme is shown in Figure 2.

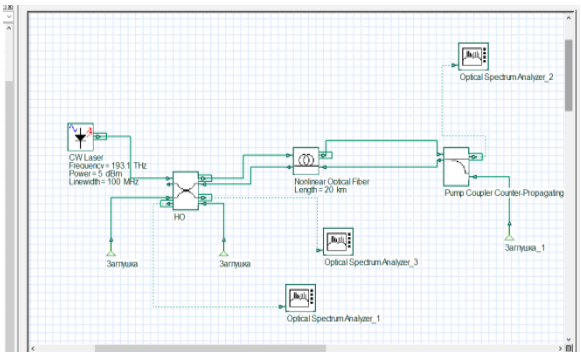


Figure 2: Simulation scheme.

The following component parameters were set in Figures 3-4.

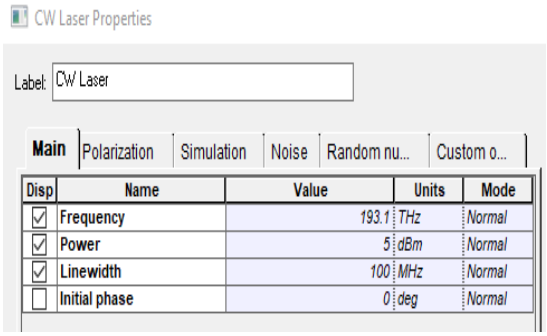


Figure 3: Laser parameters.

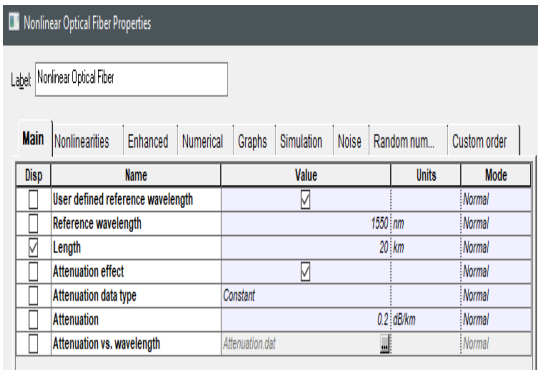


Figure 4: Optical fiber parameters.

The simulation results are shown in Figures 5-7.

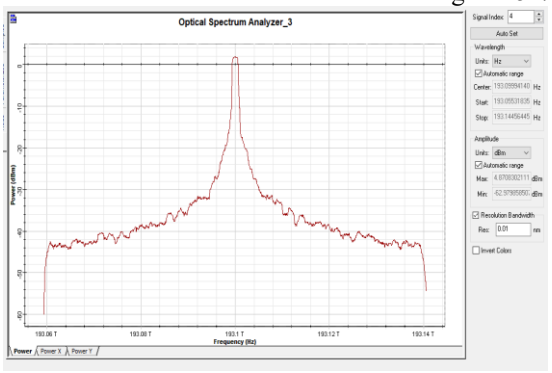


Figure 5: The signal at the output of optical fiber (Port no. 4 of the directional coupler).

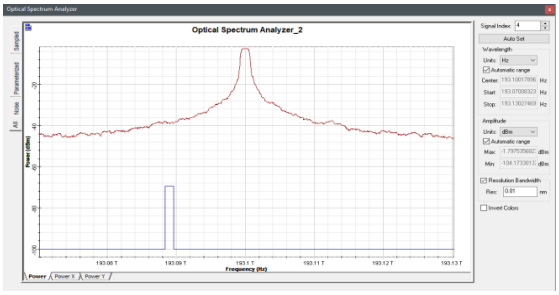


Figure 6: Signal at the output of the optical fiber.

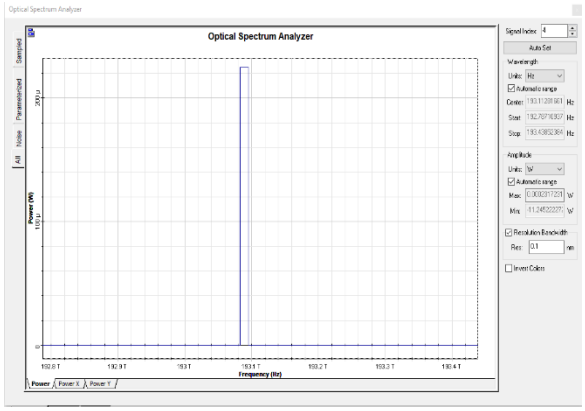


Figure 7: Spectrum of the backscattered component.

In order to test the simulation results and identify common patterns in spectrograms the experimental studies were conducted by using Brillouin optical reflectometer manufactured by the company “Ando” “AQ 8603”. For experimental studies, a one-mode optical fiber of the standard G.652 was chosen. Figures 8 and 9 show the spectrograms for the same fiber, in which a section was heated or longitudinally stretched.

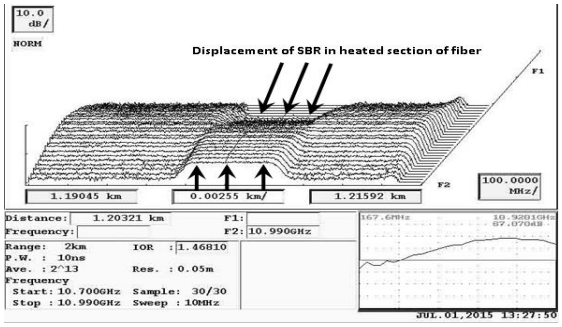


Figure 8: Spectrogram of the optical fiber section heated to 100°C.

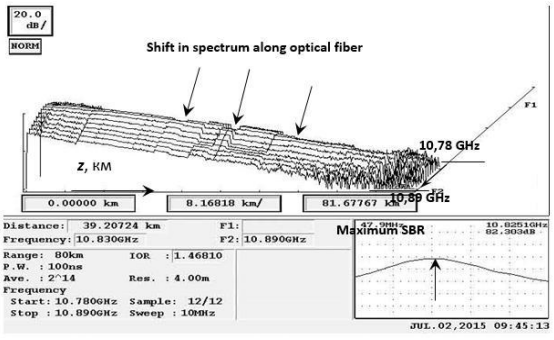


Figure 9: 3D spectrogram along OF in fiber-optic cable more than 70 km.

After spectrogram processing the picture of strain distribution in optical fiber along longitudinal coordinate is obtained. It is shown in Figure 10.

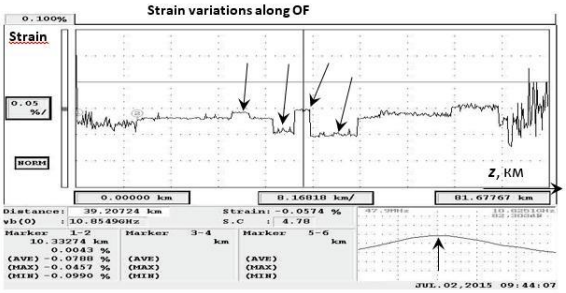


Figure 10: Strain profile in optical fiber.

Figures 8 – 10 show that changes in temperature and tension lead to a frequency shift, which makes it possible to detect areas with changes in these characteristics along the light-guide (distributed optical sensor). Each cross section in 3D-reflectogram along distance axis is a fiber reflectogram for fixed frequency. Each section along frequency axis is a profile of Brillouin spectrum in this section. In the right bottom the maximum shift of Brillouin scattering spectrum is shown. The shift equals 10.78 GHz and correlates to simulation result.

Differences in the impact factors (longitudinal tension or temperature change) in the fiber can be detected by the analysis of the back-reflected signal intensity of the Brillouin scattering ((1), Table 1). Under the stretching condition, the back-reflected signal intensity decreases. As the temperature rises, this intensity increases.

4 CONCLUSIONS

Fiber-optic sensors based on the principle of Brillouin scattering can be widely used in mining, oil and gas industries as well as in power industry, construction, aviation and space industries. They have already taken significant place in telecommunication systems for early diagnostics of damages in fiber-optic communication lines. The objectives of further research are to perform metrological analysis at all stages of the method implementation, to complete the base of Brillouin spectrograms for optical fiber of various types and to improve algorithms for automated processing spectra. The presented simulation results demonstrate the possibilities for analyzing distributed optical sensors in "OptiSystem" software.

The results obtained give evidence that Brillouin scattering spectrum analysis makes it possible to measure tension and temperature along longitudinal coordinate in optical fiber. In this process the optical fiber is both sensor and transmission media. This is the main advantage of optic-fiber sensors compared to other optic fiber sensors operating with other physical phenomena.

REFERENCES

[1] R. Bernini, A. Minardo, and L. Zeni, "Self-demodulated heterodyne frequency domain distributed Brillouin fiber sensor," IEEE Photonics Technology Letters, vol. 19, pp. 447-449, March 2007.

[2] Y. Li, X. Bao, Y. Dong, and L. Chen, "A novel distributed Brillouin sensor based on optical differential parametric amplification", Journal of Lightwave Technology, vol. 28, pp. 2621-2626, September 2010.

[3] J. Urricelqui, F. Fernandino, M. Sagues, and A. Loayssa, "Polarization diversity scheme for BOTDA sensors based on a double orthogonal pump interaction", Journal of Lightwave Technology, vol. 33, pp. 2633-2637, June 2015.

[4] M. Belal and T. P. Newson, "Experimental examination of the variation of the spontaneous Brillouin power and frequency coefficients under the combined influence of temperature and strain", Journal of Lightwave Technology, vol. 30, pp.1250-1255, May 2013.

[5] J. Luo, Y. Hao, Q.Ye, and L. Li, "Development of optical fiber sensors based on Brillouin scattering and FBG for on-line monitoring in overhead transmission lines", Journal of Lightwave Technology, vol. 31, pp.1559-1565, May 2013.

[6] F. Barrios, S. López, A. Sanz, P. Corredera, and J.Castañón, "Distributed Brillouin fiber sensor assisted by first-order Raman amplification", Journal of Lightwave Technology, vol. 28, pp.2162-2172, August 2010.

[7] H. Lee, N.i Hayashi, Y. Mizuno, and K. Nakamura, "Slope-assisted Brillouin optical correlation-domain reflectometry using polymer optical fibers with high propagation loss", Journal of Lightwave Technology, vol. 35, pp. 2306-2310, June 2017.

[8] Q. Li, J. Gan, Y. Wu, Z. Zhang, J. Li, and Z. Yang, "High spatial resolution BOTDR based on differential Brillouin spectrum technique», IEEE Photonics Technology Letters, vol. 28, pp.1493-1496, July 2016.

[9] Y.Yu, L. Luo, B. Li, K. Soga, and J.Yan, "Frequency resolution quantification of Brillouin-distributed optical fiber sensors", IEEE Photonics Technology Letters, vol. 28, pp. 2367-2369, November 2016.

[10] F. Ravet, X. Bao, Q.Yu, and L.Chen, "Criterion for subpulse-length resolution and minimum frequency shift in distributed Brillouin sensors", IEEE Potonics Technology Letters, vol. 17, pp.1504-1506, July 2005.

Comparative Analysis of File Transfer Protocols in Low-Bandwidth Radionetworks

Irina Strelkovskaya[1], Roman Zolotukhin[1] and Julia Strelkovskaya[2]
[1]*Institute of Infocommunications and Software Engineering, State University of Intellectuals Technologies and Telecommunications, Kuznechna Str. 1, 65023 Odesa, Ukraine*
[2]*Department of Criminology and Penitentiary Law, National University "Odessa Law Academy", Fontanska Str. 23, 65009 Odesa, Ukraine*
strelkovskaya@onat.edu.ua, zolotukhinrv@gmail.com, yulia.strelkovskaya@gmail.com

Keywords: UHF/VHF Radio Station, Low-Bandwidth Network, FTP, TFTP, SCP, SFTP, ETFTP, QoS Parameters.

Abstract: The use of modern digital ultra and very high frequency (UHF/VHF) radio stations in the construction of digital governmental automated control systems (ACS) of the low echelon management level has led to the creation of protocols and standards that allow data transmission in low-bandwidth communication networks. However, none of these standards provide recommendations for file transfer in communication networks facing low speed, long delay and high probability of data loss. This work investigates QoS parameters and performs comparative analysis of FTP, TFTP, SCP, SFTP, ETFTP protocols for file transfer in low-bandwidth communication networks based on UHF/VHF radio stations. A model of two Harris RF-7850M-HH radio stations connected by an attenuator and coaxial cables was used to measure the QoS parameters. The characteristics including the bandwidth, jitter and average time of data transmission depending on the operating modes of radio stations and the level of attenuation in the radio communication channel have been obtained. The time of file transfer and the actual size of the transmitted data when using these protocols was measured. The recommendations for use of file transfer protocols in radio communication channels depending on the operating mode of the radio station are given. The obtained results allow to rationally choosing the mechanism and algorithm of file transfer when building governmental ACS of the low echelon management level based on low-bandwidth communication networks to increase the efficiency of bandwidth use in radio networks.

1 INTRODUCTION

Modern governmental automated control systems (ACS) of the low echelon management level as base of command and control process are built on the basis of low-bandwidth radio networks by means of ultra and very high frequency (UHF/VHF) radio stations [1-5]. The results obtained in [6] show that the use of standard protocols for data transmission in UHF/VHF radio networks is complicated by low speed and high data delay, high jitter data delay, high probability of data loss in the channel. A number of protocols have been developed for such telecommunication channels, according to the standards STANAG 4677 [7], AdatP-36 [8], STANAG 5525 [9], etc. These standards are used to transfer information in the governmental ACS of the low echelon management level [1, 4, 5]. Modern file transfer protocols have been created for data transmission in high-speed cable networks. However, there are no recommendations or standards for transferring files in ACS built on low-bandwidth telecommunication networks. In 1996, the experimental protocol Enhanced Trivial File Transfer Protocol (ETFTP) [10] was created specifically for communication networks based on UHF/VHF radio stations with a data rate of 16 Kbps. However, today a new generation of radio stations is available with the support of data transmission of up to 1 Mbps [6]. Therefore, there is a need for a comparative analysis of existing file transfer protocols to determine the possibility of their use in low-bandwidth communication networks.

The purpose of the work is to perform a comparative analysis of file transfer protocols in low-bandwidth communication networks in terms of bandwidth efficiency of a radio communication channel. The relevance of the work is driven by the

lack of recommendations for the use of such protocols in low-bandwidth communication networks, which complicates the construction of governmental ACS of the low echelon management level on the basis of UHF/VHF radio stations.

For comparative analysis, let us consider the following file transfer protocols: File Transfer Protocol (FTP) [11], Trivial File Transfer Protocol (TFTP) [12], Secure Copy Protocol (SCP) [13], SSH File Transfer Protocol (SFTP) [14] and ETFTP [10]. FTP is one of the most common file transfer protocols. Its operation is based on the Transmission Control Protocol (TCP) [15], a client-server architecture and the ability to authenticate users to ensure secure access to data. TFTP is a simple file transfer protocol based on the User Datagram Protocol (UDP) [16] and is used mainly for the network boot of computers. It doesn`t have any authentication or encryption mechanisms. SCP and SFTP are similar in functionality to FTP, but run on Secure Shell (SSH) [17] and transmit files in the encrypted form. ETFTP is an experimental file transfer protocol designed specifically for low-bandwidth radio networks. ETFTP is configured according to the QoS parameters of the radio channel and considers the maximum possible speed and delay of data transmission. This protocol works on the basis of UDP and is focused on ensuring reliable file transfer with maximum speed and the smallest amount of service information.

Section 2 presents the results of experiments, that research and analyze the QoS parameters of file transfer and gives recommendations about their usage.

Section 3 provides the conclusions about usage of different file transmission protocols over UHF/VHF radio networks based on experiment results.

2 RESEARCH AND ANALYSIS OF QoS PARAMETERS

The scheme shown in Figure 1 was organized to take the necessary measurements. RF-7850M-HH, manufactured by Harris, are used as UHF/VHF radio stations. The radio stations are connected to each other via a coaxial cable and an attenuator with a variable attenuation level. The attenuator has three possible attenuation levels of 40, 80 and 120 dB. The power of radio stations is set to 1 W. Different levels of attenuation will simulate external interference affecting the radio channel. Personal Computers (PC) are connected to radio stations via Ethernet cable with RJ-45 (12067-5220-01) from Harris radio station accessories.

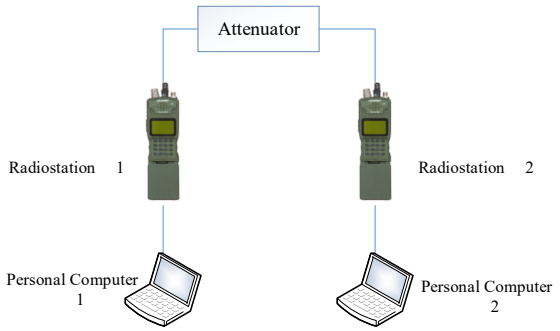


Figure 1: Scheme of the communication network for measurement.

The PC has the Ubuntu 20.04.1 LTS operating system [18]. The next utilities are used to transfer files via different protocols:

- vsftpd version 3.0.3 implements FTP, TFTP and SFTP protocols;
- OpenSSH version 1.0.2g implements SCP protocol;
- etftp and etftpd version 1.1.3 together implement ETFTP protocol.

For measurements, let us consider three modes of operation of the radio stations: FF - narrowband mode with a fixed carrier frequency, ANW2C (M-TNW) - broadband mode, QL1A - narrowband mode with frequency-hopping spread spectrum (FHSS). Let us consider 10KB, 100KB and 1MB files to be transmit via the UHF/VHF radio network. Larger files are not considered for the transmission because according to the bandwidth limits of the UHF/VHF radio network obtained in [6], such a transmission can take up to several hours, which can result in the loss of information relevance. According to [10], to estimate the effectiveness of the ETFTP protocol let us use the ratio of file size to the time of its actual transfer from one PC to another as the QoS parameter. The mentioned period includes the time to establish a connection, data transfer and disconnection. In this work, in addition, the actual size of the transmitted data is measured, which allows to evaluate the efficiency of using the bandwidth of the radio channel.

2.1 Research and Analysis of QoS Parameters in FF Mode

Three experiments were conducted to evaluate the QoS parameters and compare file transfer protocols in low-bandwidth communication networks in the

narrowband FF mode of UHF/VHF radio station. Let us consider three cases of measuring the QoS parameters at attenuation of 40, 80 and 120 dB, respectively.

The first experiment: set the attenuator to attenuation of 40 dB, and the radio station to the FF mode. Similarly to [1], the measurements of QoS parameters of the low-bandwidth radio network were performed and the following values were obtained:

- bandwidth - 102 kbit/s;
- jitter - 126 ms;
- the average ping – 1189 ms.

Files of different sizes from PC №2 to PC №1 were transferred, and measured the time of file transfer and the actual size of the transmitted data was measured. The results are shown in Table 1.

Table 1: The measurement result in the FF mode, 40 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	14	13	22	112	81	1149
TFTP	62	14	660	142	6840	1523
SFTP	10	12	27	113	115	1249
SCP	13	14	41	115	115	1154
ETFTP	23	11	45	109	298	1141

The second experiment: set the attenuator to attenuation of 80 dB. The research showed that QoS parameters have the following values:

- bandwidth - 100 kbit/s;
- jitter - 128 ms;
- the average ping – 1239 ms.

The results of file transfer measurements are shown in Table 2.

Table 2: The measurement result in the FF mode, 80 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	20	12	40	112	120	1145
TFTP	63	12	662	115	6933	1517
SFTP	22	12	41	113	120	1131
SCP	38	12	66	115	249	1156
ETFTP	28	11	42	111	256	1185

The third experiment: set the attenuator to attenuation of 120 dB. The research showed that QoS parameters have the following values:

- bandwidth – 58,8 kbit/s;
- jitter – 154 ms;
- the average ping – 1297 ms.

The results of file transfer measurements are shown in Table 3.

Table 3: The measurement result in the FF mode, 120 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	10	13	30	112	88	1164
TFTP	44	13	720	113	7312	1589
SFTP	24	12	39	113	145	1131
SCP	60	14	103	111	240	1154
ETFTP	21	12	48	110	360	1185

After analyzing the results shown in Table 1, Table 2 and Table 3, it is not difficult to see that in the FF mode, the transfer of 10 KB files using FTP was from 10 to 20 s, 100 KB files - from 22 to 40 s, and 1 MB files - from 81 to 120 seconds. TFTP transmitted 10 KB files from 44 to 66 s, 100 KB files from 660 to 720 s, and 1 MB files from 6840 to 7312 s. ETFTP, SFTP, SCP in this case show worse data transfer rates than FTP, but better in contrast to TFTP. The ratio between the file size and the actual amount of transferred data files of 10 KB, 100KB, 1MB for FTP, SFTP, SCP, ETFTP protocols ranges from 10% to 40%. However, for the TFTP protocol, this ratio is from 20% to 59%. Thus, in the narrowband FF mode it is rational to use the FTP protocol.

2.2 Research and Analysis of QoS Parameters in QL1A Mode

Three experiments were conducted to evaluate QoS parameters and compare file transfer protocols in low-bandwidth communication networks in the narrowband QL1A mode of the UHF/VHF radio station. Let us consider the three cases of measuring QoS parameters at attenuation of 40, 80 and 120 dB, respectively.

The first experiment: set the attenuator to attenuation of 40 dB, and the radio station in the QL1A mode. The research showed that QoS parameters are significantly different from the FF mode and have the following values:

- bandwidth – 17,4 kbit/s;
- jitter – 453 ms;
- the average ping – 3969 ms.

Files of different sizes from PC 2 to PC 1 were transferred according to previous measurements. The results of measurements are shown in the Table 4.

Table 4: The measurement result in the QL1A mode, 40 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	57	13	372	112	3757	1621
TFTP	267	16	2321	169	26743	1723
SFTP	56	11	195	113	2242	1649
SCP	53	15	240	120	2755	1654
ETFTP	51	11	165	110	1891	1579

The second experiment: set the attenuator to attenuation of 80 dB. The research showed that QoS parameters have the following values:

- bandwidth – 15,7 kbit/s;
- jitter – 476 ms;
- the average ping – 4137 ms.

The results of file transfer measurements are shown in the Table 5.

Table 5: The measurement result in the QL1A mode, 80 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	65	13	363	113	3832	1626
TFTP	308	12	2417	169	27631	1763
SFTP	72	11	256	113	2312	1659
SCP	78	12	480	121	2933	1674
ETFTP	65	12	240	114	1911	1583

The third experiment: set the attenuator to attenuation of 120 dB. The research showed that QoS parameters have the following values:

- bandwidth – 12,5 kbit/s;
- jitter – 489 ms;
- the average ping – 4279 ms.

The results of file transfer measurements are shown in Table 6.

After analyzing the results shown in the Table 4, Table 5 and Table 6, it is not difficult to see that in QL1A mode ETFTP transmits 10 KB files for 51 – 66 s, 100 KB for 160 - 253 s, and 1 MB for 1891 – 1974 s. The ratio between the file size and the actual amount of data transmitted 10 KB, 100 KB, 1 MB files for ETFTP ranges from 10% to 61%. Other protocols have worse time and actual data rates. Thus, the ETFTP protocol provides the fastest transfer of files and minimal data overhead in the QL1A mode.

Table 6: The measurement result in the QL1A mode, 120 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	120	13	346	113	3951	1723
TFTP	335	40	720	113	29767	1871
SFTP	68	12	340	113	2432	1789
SCP	70	15	293	111	3123	1811
ETFTP	66	12	253	116	1974	1613

2.3 Research and Analysis of QoS Parameters in ANW2C Mode

Three experiments were conducted to evaluate QoS parameters and compare file transfer protocols in low-bandwidth communication networks in the narrowband ANW2C mode of the UHF/VHF radio station. Let us consider the three cases of measuring QoS parameters at attenuation of 40, 80 and 120 dB, respectively.

The first experiment: set the attenuator to attenuation of 40 dB, and the radio station in the ANW2C mode. The study showed that this mode of operation has the best QoS parameters compared to other modes:

- bandwidth – 259 kbit/s;
- jitter – 14 ms;
- the average ping – 273 ms.

Files of different sizes from PC №2 to PC №1 were transferred according to the previous measurements. The results are shown in the Table 7.

Table 7: The measurement result in the ANW2C mode, 40 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	1	12	3	112	34	1143
TFTP	5	12	62	123	571	1265
SFTP	1	12	4	113	33	1151
SCP	1	14	3	119	34	1157
ETFTP	7	11	11	109	51	1124

The second experiment: set the attenuator to attenuation of 80 dB. The research showed that QoS parameters have the following values:

- bandwidth – 201 kbit/s;
- jitter – 15 ms;
- the average ping – 319 ms.

The results of file transfer measurements are shown in Table 8.

Table 8: The measurement result in the ANW2C mode, 80 dB.

	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	1	12	4	112	33	1178
TFTP	7	12	60	118	555	1102
SFTP	1	12	4	113	33	1150
SCP	1	14	5	114	60	1156
ETFTP	3	12	21	116	83	1185

The third experiment: set the attenuator to attenuation of 120 dB. The research showed that QoS parameters have the following values:

- bandwidth – 150,8 kbit/s;
- jitter – 34 ms;
- the average ping – 327 ms.

The results of file transfer measurements are shown in Table 9.

Table 9: The measurement result in the ANW2C mode, 120 dB.

.	10 KB file		100 KB file		1MB file	
	Time, s	Data size, KB	Time, s	Data size, KB	Time, s	Data size, KB
FTP	1	12	7	112	68	1143
TFTP	6	12	55	111	555	1142
SFTP	1	12	4	113	35	1150
SCP	1	12	5	112	67	1102
ETFTP	5	11	22	116	240	1185

It is not difficult to see that according to the results shown in Table 7, Table 8 and Table 9, in the ANW2C mode, the FTP, SFTP and SCP protocols transmit 10 KB files in 1 s, 100 KB in 3 - 7 s, and 1 MB for 33 - 68 s. TFTP with ETFTP transferred 10 KB files in 3 - 7 s, 100 KB in 11 - 62 s, 1 MB in 51 - 571 s. The ratio between file size and the actual amount of data transferred was 10% to 27% for all protocols. Thus, in this mode it is rational to use protocols FTP, SCP, SFTP.

2.4 Recommendations for Use of File Transfer Protocols in Low-Bandwidths Networks

Based on the experiments, the following recommendations can be proposed:

1) It is advisable to use ETFTP in narrowband, low latency and low bandwidth radio networks.

2) In broadband radio channels with low latency and high bandwidth, compared to narrowband radio modes, it is advisable to use FTP protocol.

3) TFTP has the worst QoS parameters when transferring files, but it should be used when you need to occupy only part of the maximum bandwidth.

4) In the narrowband mode of UHF/VHF radio stations, it is rational to compress files before sending them to reduce data transmission time.

The application of the research results allows to rationally use the existing file transfer protocols in the construction of governmental ACS of the low echelon management level depending on the type and conditions of use of modern UHF/VHF radio stations.

3 CONCLUSIONS

1) The QoS parameters of FTP, TFTP, SCP, SFTP and ETFTP file transfer protocols in low-bandwidth radio networks based on UHF/VHF radio stations were measured.

2) The analysis of expediency for various file transfer protocols is carried out. Recommendations for their use depending on the operating modes of UHF/VHF radio stations are given.

3) It is not advisable to transfer files larger than 1 MB in narrowband modes of radio stations because a significant delay in data transmission can result in losses of information relevance.

4) Approbation of the research results allows building governmental ACS of the low echelon management level with the ability to transfer files in low-bandwidth networks with the rational use of bandwidth of the radio channels.

5) The research results of this article and publications [1, 6] represent the part of work concerning development of telecommunication system model with QoS parameters in UHF/VHF radio networks.

REFERENCES

[1] I. V. Strelkovskaya, R. V. Zolotukhin, and A. O. Makoganiuk, "Modeling of telecommunication components of automated control systems in low-bandwidth radio networks" / Springer Science & Business Media, 2021.

[2] J. S. Bayne, "A Theory of Enterprise Command and Control", MILCOM 2006 - 2006 IEEE Military Communications conference, 23-25 Oct. 2006, Washington, DC, USA, doi: https://doi.org/10.1109/MILCOM.2006.302294.

[3] J. Lawson, "Command control as a process", IEEE Control Systems Magazine, pp. 5-11, March 1981, doi: https://doi.org/10.1109/MCS.1981.1100748.

[4] R. Masnica and J. Štulrajter, "Development of Interoperability C4IS", 7th International Scientific Conference Communication and Information Technologies, 9 – 11 Oct. 2013, Starý Smokovec, Slovakia, ISBN 978-80-8040-464-2.

[5] J. Rhea, "Seamless communications: the challenge of tactical command and control", Military&AeroSpace Electronics, Jan 1st, 1997, [Online]. Available: https://www.militaryaerospace.com/communications/article/16710327/seamless-communications-the-challenge-of-tactical-command-and-control, accessed March 2021.

[6] I. Strelkovskaya and R. Zolotukhin, "Research of low-bandwidth radionetworks QoS parameters" //Information and Telecommunication Sciences, International Research Journal, Volume 11, Number 1(20), January-June 2020, doi: https://doi.org/10.20535/2411-2976.12020.77-81.

[7] STANAG 4677: 2014 Dismounted soldier systems standards and protocols for command, control, communications and computers (C4) interoperability./ NATO 2014.

[8] NATO - ADATP-36: Friendly force tracking systems (FFTS) interoperability / NATO 2017.

[9] STANAG 5525: 2007 Joint C3 Information Exchange Data Model - JC3IEDM/ NATO 2007.

[10] Experiments with a Simple File Transfer Protocol for Radio Links using Enhanced Trivial File Transfer Protocol (ETFTP) [Online]. Available: https://tools.ietf.org/html/rfc1986, December 2020.

[11] File transfer protocol (FTP), RFC 768 [Online]. Available: https://tools.ietf.org/html/rfc768, December 2020.

[12] The TFTP protocol (revision 2), RFC 1350 [Online]. Available: https://tools.ietf.org/html/rfc768, December 2020.

[13] "Linux and Unix scp command". Computer Hope [Online]. Available: https://www.computerhope.com/unix/scp.htm, December 2020.

[14] SFTP - SSH Secure File Transfer Protocol [Online]. Available: https://www.ssh.com/ssh/sftp/, December 2020.

[15] Transmission Control Protocol, RFC 793 [Online]. Available: https://tools.ietf.org/html/rfc793, December 2020.

[16] User Datagram Protocol, RFC 768 [Online]. Available: https://tools.ietf.org/html/rfc768, December 2020.

[17] The Secure Shell (SSH) Transport Layer Protocol, RFC 4253 [Online]. Available: https://tools.ietf.org/html/rfc4253, December 2020.

[18] Ubuntu 20.04.1 LTS (Focal Fossa) [Online]. Available: https://releases.ubuntu.com/20.04/, December 2020.

Mathematical Description of Control Problems in SDN Networks

Oleksandr Romanov[1], Eduard Siemens[2], Mikola Nesterenko[1] and Volodymyr Mankivskyi[1]
[1]*Institute of Telecommunication Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Peremohy avenue 37, 03056 Kyiv, Ukraine*
[2]*Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*
a_i_romanov@ukr.net, eduard.siemens@hs-anhalt.de, nikolaiy.nesterenko@gmail.com, v.b.mankivskiy@gmail.com

Keywords: Controller, SDN, Software-Defined Networking, Open Network Operating System, ONOS, Control Plane Disaggregation, Message Distribution System, OpenFlow, Datapath Model.

Abstract: To improve the efficiency of networks, fully automate the processes of network management, administration and technological, reduce the time to market for modern services, ensure the introduction of new technologies throughout the network, and not gradually introduce them at individual sites, operators are starting the practical use of software-defined networks with virtualization network functions. The optimality of the decisions made, largely depends on the capabilities of the control plane in the SDN architecture. To obtain optimal solutions, a mathematical formulation of control problems is required, using real indicators of the functioning of telecommunication networks as input data. The paper proposes mathematical models of the process of functioning of the SDN network in the main modes of operation: in the mode of planned changes in the structure of the network, in the mode of normal functioning, when all performance indicators are within the specified values, in the overload mode. The results obtained make it possible to determine the list of required network elements, to distribute functions between the elements of the control system, to develop requirements for the monitoring system to collect the necessary initial data and to determine the interaction algorithm of the elements to obtain optimal solutions in the process of network management.

1 INTRODUCTION

Traditional telecommunication networks represent a set of specialized physical devices such as routers, switches, soft switches, firewalls, and other equipment. These devices were created based on specific hardware and software platforms from different vendors. Therefore, the implementation of new modern services on networks, as a rule, requires replacing the old equipment set with a new one. This approach leads to the occurring of long design cycles, purchase of the necessary equipment and conducting commissioning tasks. All this negatively works on the efficiency of providing users with new products and services. Maintaining and control such a network control is ineffective and expensive costs. Therefore, quite often, investments in the development of the network to satisfy subscribers' requests could significantly exceed the growth in income from the services` provision.

However, nowadays the networks of telecommunications operators are mainly consisting of "monolithic physical" network elements, where control functions, administration and transmission of user data are performed by physical devices. Often the network is built using network elements from the same producers, as this makes it easier to ensure compatibility. Deployment of services, modification ("upgrades") of equipment or services are performed by tuning on each network element and requires close coordination of internal and external operator's resources. This approach makes the operator's network inflexible, makes it difficult to implement new services and functions, and also increases the operator's dependence on specific vendor solutions.

Therefore, now, the issues of building networks based on the SDN concept are increasingly on the agenda of research organizations, universities and mobile operators. Representatives of the Open Networking Foundation (ONF) consortium make hereby the main contribution to the development of this area. ONF is an non-profit organization dedicated to advancing of SDN and introduction of NVF [1].

For the moment, ONF has developed a number of documents describing the principles of construction and performance of SDN networks. In [3, 4, 5, 6] general requirements, system approaches and

and performance of SDN networks. In [3, 4, 5, 6] general requirements, system approaches and generalized architecture of SDN networks are reviewed. In works [2, 7, 8, 11] tasks and features of the protocols used in solving various problems in SDN networks are considered. Ins [9, 10, 12] the features of the construction of SDN network elements and the order of their interaction in the process of maintaining traffic flows are described. In [3, 4, 5, 6, 14] the principles of constructing an optical transport network and are given recommendations to provide the safety of their operation are considered. The most complete and systematized material is presented in [8, 9]. In articles [10, 11, 12, 13] various aspects of servicing information flows are investigated and focused on the need to meet the requirements for providing network security.

It should be noted that most of the research work focus on the ways of practical implementation of SDN principles during network deployment. Also considered are possible structure and order of interaction of the main elements of the network, such as a controller and a switch. At the same time, one of the most difficult and important issues in SDN networks is to provide the solution of managing problems under different operating conditions. The solution of these tasks can be assigned both to the controller or network operating system, and to control applications located in the control plane. In this research, possible formulations of control problems and their formalized representation as mathematical models (equations) shall be analyzed.

2 SDN MANAGEMENT TASKS

Native existing telecommunications networks based on "monolithic physical" elements are forcing an increasing number of telco operators to take part in the development of the ONF consortium. They develop experimental areas of the network in which studies of the practical availability of any given solution are carried out.

2.1 Analyses of SDN Requirements

Let's consider what requirements telecom operators impose on networks based on SDN technology:

Improve operational efficiency:

- provide flexibility and scalability of the entire operator's network;
- fully automate the processes of performing operations, administration (Management) and maintenance, OAM;
- solve the problems of dynamic control of traffic flows in real time, in accordance with the current state of network resources;
- operatively create the necessary types of services what require the combined usage of several services.

Transform business models:

- reduce the output time of modern services to market;
- provide that innovations are quickly extended across the entire network, rather than being set in across the network sections;
- quickly and efficiently create and provide demanded services (Agile);
- increase the quality of the process of providing services to end users.

All this is achieved through the implementation of new approaches to the building of SDN networks, specifically:

- separating the control plane from the data plane;
- providing programmability of the level of control of network resources, computing resources, resources data storage and service orchestration;
- providing the ability to virtualize most types of equipment and system network functions;
- implementation of providing users with the same set of services, regardless of whether a set of physical devices is used for this or their representation in the form of virtual machines;
- unification of protocols and automation of solving problems of network elements` configuration;
- providing an individual mechanism for administration and allocation of resources in the network upon request of various services and functions;
- automation of management processes in the deployment of network elements and business processes.

Assurance the conditions for meeting these requirements will allow creating telecommunication networks with high efficiency and competitive ability at the telecommunications market.

2.2 Components of SDN Architecture

Let's analyze the components of the SDN network architecture, which are defined in ONF TR-502 [3]. At the same time, focus will be on the elements that are directly involved in the management process. As far

as, at the present stage, SDN networks will operate in the environment of traditional networks, we will consider an architecture that takes this factor into account (Figure 1). We will also take into account the possibility of interference in the control process of a human operator.

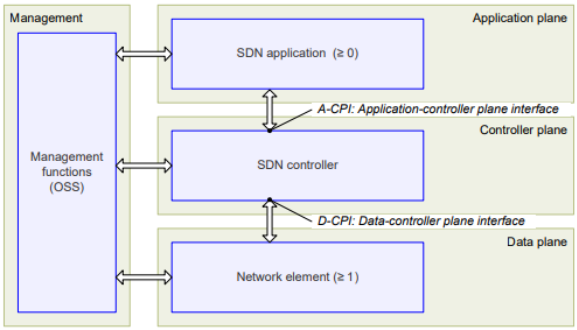


Figure 1: SDN with function management.

At Figure 1 are showed the following main components of SDN:

- data plane;
- controller plane;
- application plane;
- management plane.

The data plane consists of one or more network elements. Each element contains a set of resources for forwarding or processing traffic. Resources are abstractions of the basic physical capabilities of an element. The data plane interacts with the control plane through southbound interfaces, often called as SBI.

The application plane can contain one or more applications. Each application has exclusive control over the set of resources provided by one or more SDN controllers. Applications can:

- interact with each other directly;
- act as SDN controller;
- place their network requirements on the controller plane through northbound interfaces, often called NBI.

The control plane consists of a set of SDN controllers. Each SDN controller manages a set of resources for one or more network elements in the data plane. The minimum functionality of an SDN controller should provide:

- processing application requests that are assigned to it;
- isolate applications' work from each other. For this the SDN controller can communicate with other SDN peer-to-peer controllers;
- respond to network accidents to restore normal operation after a failure;
- manage the competing requirements of different applications.

An independent element of the architecture is the management functional block. This block provides for the possibility of intervention in the control process of a human operator. It can be used:

- for initial configuration of network elements;
- when assigning areas controlled by different SDN controllers;
- to configure the SDN area controller.

At the controller level, the human operator can set up:

- policies that define the scope of control that are provided to SDN applications;
- limits of allocated resource;
- list of system parameters to be monitored.

At the application level, the human operator usually configures:

- contracts and service level agreements (SLA);
- algorithms for solving control problems;
- device priority level when solving problems together.

At all levels of control, human-operator sets up a security policy that allows distributed functions to communicate safely with each other.

2.3 Components of SDN Architecture

To solve management issues, it is necessary to have elements that collect information, analyze it, make decisions and communicate commands to executive agents. These elements in the SDN architecture are agents and coordinators.

At Figure 2 are showed the agents and coordinators in the SDN controller and network elements.

Agents support the concept of sharing (co-use) or virtualizing default (essential) resources. For example, agents own next information:

- which network elements ports are monitored by SDN;
- which elements of the virtual network are open for SDN applications;
- how to isolate the service of one client from another one.

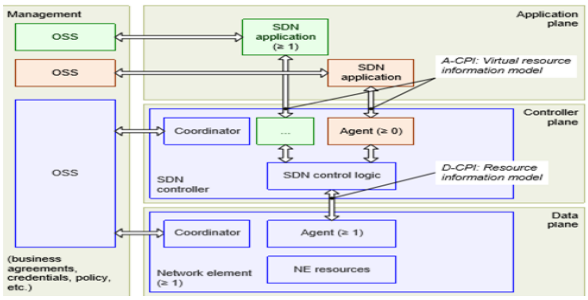


Figure 2: Agents and Coordinators in SDN Architecture.

In the SDN controller, different agents can control the network at different levels of abstraction and provide the execution of different set of functions. The purpose of the SDN control logic is to:

- provide arbitration between the network requirements of all SDN applications;
- develop a set of instructions for a network element (NE) and implement them through NE agents.

Coordinators in the NE and SDN controller establish specific resources and policies to customer from the human operator. Multiple agents can exist on any network element and SDN controller. But there is only one logical control interface. Therefore, at any time, only one coordinator works per one network element or an SDN controller.

3 FORMALIZED STATEMENT OF CONTROL ISSUES FORMED ON THE CONTROLLER PLANE

The functioning of SDN networks requires constant management of its elements. In this case, the nature of the tasks to be solved can be considered in different conditions as:

- planned deployment or deployment of SDN network;
- normal operation mode of the SDN network;
- functioning of SDN networks in extreme situations (case).

For each of the above-mentioned conditions can be earmarked specific groups of control tasks. So, for the first case, the group of management tasks with a planning-changing structure should include:

- analysis of the plan of deployment, rolling up or reorganization of the network;
- analysis of an existing, or required (when deploying), or becoming free (rolling up) network resource;
- analysis of requirements for quality of service QoS and network bandwidth;
- formation of a load distribution plan (LDP);
- making decisions on the distribution of the network resource, providing the implementation of the LDP with minimal time, material-technical and other types of costs;
- formation of teams for network elements that implement the decisions.

The tasks of this group can be solved in the controller plane or assigned to applications. It should be noted that one of the main tasks of network planning is to determine the required performance of the branches and the channel resource that provide service of the incoming load with a desired quality.

This task can be formulated like this. Determine the required performance of each branch of the network, expressed in the number of standard channels $\{V_{ij}\}$, in the network given by the graph G(N,M), and formulate the LDP, which represents a matrix $M_v = \{\mu^v_{ij}\}$ set of paths μ^v_{ij} ($v = \overline{1,k}$) transferring information between each pair of Switching center (SC) i,j ($i,j = \overline{1,N}$, where N is a number of SC in the network), with order v ($v = \overline{1,k}$) their occupations, providing service to the incoming load the network Z by I information directions J_{ij} ($i,j = \overline{1,N}$) with a given quality of service P_{ij} with the minimum total involved channel capacity V_Σ of the branches m_{ij} ($i,j = \overline{1,N}$), i.e.

$$\{V_{ij}\} = f(M_v = \{\mu^v_{ij}\}),\ (i,j = \overline{1,N},\ v = \overline{1,k}) \quad (1)$$

under the using restrictions:

$$\begin{cases} V_\Sigma = \min \sum_{i=1}^{N} \sum_{j=1}^{N} V_{m_{ij}} \\ P \geq \max\{P_{ij}\}, (i,j = \overline{1,N}) \\ P_{ij} = \prod_{i=1}^{N} \left[1 - \prod_{j=1}^{N} \left(1 - p_{m_{ij}}\right) \right] \\ Z = \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ij}, (i,j = \overline{1,N}) \\ k = \max \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{N} k_{ij}, (i,j = \overline{1,N}) \\ N, M, V_{ij}, k_{ij} \in G(N,M) \\ I = N(N-1)\ \ at\ J_{ij} \neq J_{ji} \end{cases}$$

where P - permissible denial-of-service rate in the direction of communication;

P_{ij} – real probability of fault of service in information directions ($i,j = \overline{1,N}$);

k_{ij} - utilization factor in branches m_{ij} ($i,j = \overline{1,N}$);

z_{ij} – incoming load to the service in information directions;

$p_{m_{ij}}$ - probability of losses on branches;

M - number of branches in the network.

Equation (1) is rather complicated and solutions cannot always be obtained. Therefore, sometimes solve a little simplified problem:

$$\mathrm{V}=\left\{V_{ij}\right\},\left(i,j=\overline{1,N}\right) \qquad (2)$$

using the following system of restrictions:

$$\begin{cases} V_{\Sigma}=\min\sum_{i=1}^{N}\sum_{j=1}^{N}V_{m_{ij}} \\ M_{\nu}=\left\{\mu_{ij}^{\nu}\right\},\left(i,j=\overline{1,N},\nu=\overline{1,k}\right) \\ P\geq\max\left\{P_{ij}\right\},(i,j=\overline{1,N}) \\ P_{ij}=\prod_{i=1}^{N}\left[1-\prod_{j=1}^{N}\left(1-p_{m_{ij}}\right)\right] \\ Z=\sum_{i=1}^{N}\sum_{j=1}^{N}z_{ij},(i,j=\overline{1,N}) \\ k=\max\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{N}k_{ij},(i,j=\overline{1,N}) \\ N,M,V_{ij},k_{ij}\in G(N,M) \\ I=N(N-1)\ \ at\ J_{ij}\neq J_{ji} \end{cases}$$

That is, in the process of calculating the required channel capacity of the network branches, it is assumed that the LDP is given.

During the normal functioning of the network, the main management tasks are:

- monitoring the technical condition of network elements and estimation the values of the parameters of its functioning;
- method, recording and processing of data on the technical failures in network elements, leading to a degradation in the parameters of the functioning of communication directions;
- control of indicators of bandwidth of branches and directions of communication;
- establishing the facts of deviation of parameter values from the established norm;
- identification of places and causes of network failure;
- determining a way to restore the normal network functioning;
- making a decision to restore the normal network functioning;
- formation of a team for control objects in order to implement the decision.

The main task of the SDN network controller is to form a LDP (load distribution plan) that provides the specified characteristics and efficiency of the network equipment use, i.e.:

$$\mathrm{M}_{\nu}=\{\mu^{\nu}_{ij}\}\ (i,j=\overline{1,N},\ \nu=\overline{1,k}) \qquad (3)$$

under the following system of restrictions

$$\begin{cases} V^{*}=\max\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left[f\left(P_{ij}\right)\left(V_{ij}-V_{ij}^{0}\right)\right]\right\} \\ V=\left\{V_{m_{ij}}\right\},(i,j=\overline{1,N}) \\ Z=\left\{Z_{m_{ij}}\right\},(i,j=\overline{1,N}) \\ P\geq\left\{P_{ij}\right\},(i,j=\overline{1,N}) \\ P_{ij}=\prod_{i=1}^{N}\left[1-\prod_{j=1}^{N}\left(1-p_{m_{ij}}\right)\right] \\ V_{ij},Z_{ij},P_{ij},p_{m_{ij}},N,M\in G(N,M) \end{cases}$$

where V^{*} - the number of channels, deleting which from the branch does not lead to $P_{ij} > P$;

$$f(P_{ij})=\begin{cases}1,if\ \max\{P_{ij}\}\leq P\\ \gamma_{ij},if\ \max\{P_{ij}\}>P\end{cases};$$

$$\gamma_{ij}=\begin{cases}1,or\ V_{ij}\geq V_{ij}^{0}\\ 0,or\ V_{ij}<V_{ij}^{0}\end{cases};$$

V_{ij} - number of channels in a branch;

V^{0}_{ij} - required number of channels in a branch m_{ij} to service the incoming load with a given quality.

The value of V^{*} can acts as an assessment of management efficiency when comparing different SDN management applications.

If the LDP generated by the SDN controller is close to optimal, then in the normal functioning of the communication network, as a rule, the operational-technical requirements will be satisfied.

An extreme situation may turn out to be one of the conditions for the “normal” functioning of a controlled communications network. Under such conditions, the role of SUSS is especially significant, and the tasks solved by this system have a number of specific features. These tasks, first of all, can be attributed:

- organization of a system for collecting data about damage on the elements and sections of a controlled communication network;

- method and processing of data about the damage on the elements and sections of the network;
- determining the type of damage to elements and sections of the network;
- accounting of the resource of the managed network, providing the functioning of the existing communication directions;
- accounting for the reserve network resource and determining the places of its exploitation;
- estimation of the expediency of introducing restrictions on incoming load;
- formation (re-formation) of LDP for the existing directions of communication;
- clarification of requirements for the quality of service in the areas of communication;
- making decisions for the restoration of damaged network elements, reducing the influence of damaging factors;
- formation and transmission of commands to control objects and providing control over their implementation.

In the event of extreme situations on the communication network associated with the impact of external damaging factors on it, the Controller is transferred to a special operating mode. To estimate the situation on the Controller communication network, using (3) will be solved. As a result, it may turn out that the quality of service $P < \max\{P_{ij}\}$ in some directions of communication below the required standards. Under these conditions, first of all, an attempt is made to bring the characteristics of the network to the required standards by redistributing and limiting the number of ways of transmitting information in terms of communication directions, i.e.

$$\nu = \min \sum_{i=1}^{N}\sum_{j=1}^{N}\left[\nu_{ij} - \nu'_{ij}\right]\left(i, j = \overline{1,N}\right) \qquad (4)$$

under the following restrictions:

$$\begin{cases} M_\nu = \left\{\mu_{ij}^{\nu}\right\}, \left(i, j = \overline{1,N}, \nu = \overline{1,k}\right) \\ V = \left\{V_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ Z = \left\{Z_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ P \geq \max\left\{P_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ N, M, Z, P, V, M_\nu \in G\left(N, M\right) \end{cases},$$

where ν_{ij} - possible ways of transferring information in the direction,

ν'_{ij} ways of transmitting information, which are prohibited from usage.

If the measures taken do not lead to the desired result, then the next step is to put into operation the existing reserve of forces and means. At this stage, the minimum required additional resource is determined, which provides that the network characteristics are brought to the required standards, i.e.

$$V_0 = \min\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left[m_{ij}\left(V_{ij} - V_{ij}^{0}\right)\right]\right\}\left(i, j = \overline{1,N}\right) \qquad (5)$$

with the following restrictions:

$$\begin{cases} M_\nu = \left\{\mu_{ij}^{\nu}\right\}, \left(i, j = \overline{1,N}, \nu = \overline{1,k}\right) \\ Z = \left\{Z_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ P \geq \max\left\{P_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ N, M, Z, P, V, M_\nu \in G\left(N, M\right) \end{cases},$$

$$m_{ij} = \begin{cases} 0, if\ V_{ij} \geq V_{ij}^{0} \\ 1, if\ V_{ij} < V_{ij}^{0} \end{cases},$$

where V^{0}_{ij} – the actual (available) resource of channels in the network;

V^{0}_{ij} - the resource of channels in the network, which is necessary to ensure a given quality of service;

V_0 - the number of channels that must be added to the network in order to bring the characteristics to the required standards.

If the required number of channels exceeds the available reserve, then the next measure to bring the network characteristics to the required standards may be to limit the incoming load, i.e.

$$Z_0 = \min \sum_{i=1}^{N}\sum_{j=1}^{N}\left(Z_{ij} - \Delta Z_{ij}\right)\left(i, j = \overline{1,N}\right) \qquad (6)$$

under the following restrictions:

$$\begin{cases} M_\nu = \left\{\mu_{ij}^{\nu}\right\}, \left(i, j = \overline{1,N}, \nu = \overline{1,k}\right) \\ V = \left\{V_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ P \geq \max\left\{P_{ij}\right\}, \left(i, j = \overline{1,N}\right) \\ N, M, P, V, M_\nu \in G\left(N, M\right) \end{cases},$$

It should be noted that it is rather difficult to obtain a solution to control problems (4), (5), (6) for real large-scale switched networks.

Sometimes, in the event of extreme situations, instead of the requirement to bring the network characteristics to the normalized values, a condition is set to provide the possible bandwidth with the maximum acceptable quality. In this case, the specified quality of service is provided for customers and messages of higher categories. At the same time, the quality of service for the lower categories is not standardized.

The availability of a large number of random factors affecting the conditions for the functioning of SDN networks leads to the advisability of introducing integrated management of the structure, flows, parameters and modes of network operation in order to maintain their basic characteristics within the required norms.

4 CONCLUSIONS

In this work, an analysis of the functioning process of the control (managment) level of the SDN network was conducted. Hereby, the purpose and functions of the list of the main elements were conducted. Also, the process of interaction of system elements in solving arising problems was analyzed.

The main attention is paid to the determination of the list of tasks for SDN network management in various modes of exploitation. A possible variant of a group of practical control problems, which might arise in real networks was conducted.

An important work aspect of this consideration is the formalization of management tasks. For this, three most widespread modes of network functioning, were considered, along with possible methods for solving control problems. For the latter ones, their mathematical description was given. The advantage of mathematical models for solving control problems is that the values of real indicators of network functioning are used as input data and parameters.

REFERENCES

[1] Open Network Foundation. Accelerating the Adoption of SDN & NFV, 2021.

[2] Open Networking Foundation. OpenFlow Switch Specification Version 1.5.1 (Protocol version 0x06), 2015.

[3] K. Pentikousis, IETF RFC 7426. Request for Comments: 7426. ISSN: 2070-1721 EICT.

[4] O. I. Romanov, M. V. Oryschuk, and Y. S. Hordashnyk. “Computing of influence of stimulated Raman scattering in DWDM telecommunication systems”, UkrMiCo, pp. 199-209, 2016.

[5] L. Globa, M. Skulysh, O. Romanov, and M. Nesterenko, “Quality Control for Mobile Communication Management Services in Hybrid Environment”, UkrMiCo, pp. 133-149, 2018.

[6] K. Pentikousis, ONOS. Security and Performance. Analysis: Report No. 1. September 19, 2017.

[7] O. Romanov and V. Mankivskyi, “Optimal Traffic Distribution Based on the Sectoral Model of Loading Network Elements”. [2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T)], 2019.

[8] O. Romanov, M. Nesterenko, and L. Veres, “Methods for calculating the performance indicators of ip multimedia subsystem (IMS)” in Lecture Notes in Networks and Systems, 2021, pp. 229-256.

[9] C. C. O'Connor, T. Vachuska, and B. Davie, “Software-Defined Networks: A Systems Approach”, 2021, p. 152.

[10] K. Phemius, M. Bouet, and J. Leguay, “Distributed Multi-domain SDN Controllers” in Thales Communications & Security, 2013, pp.198-209.

[11] J. Lam, S. Lee, and O. Yustus, “Securing SDN Southbound and Data Plane Communication with IBC” in Hindawi Publishing Corporation Mobile Information Systems, Volume 2016, p.12.

[12] K. Phemius, M. Bouet, and J. Leguay, “ONOS Intent Monitor and Reroute service: enabling plug&play routing logic” in Thales Communications & Security, 2013, p. 19.

[13] D. Comer and A. Rastegarnia, “Externalization of Packet Processing in Software Defined Networking”, 2019, p. 22.

[14] O. Romanov, M. Nesterenko, and V. Mankivskyi, “The usage of regress model coefficient utilization of channels for creating the load distribution plan in network” in visnyk ntuu kpi seriia-radiotekhnika radioaparatobuduvannia, 2016, pp. 34-42.

Multilevel Ontologies for Big Data Analysis and Processing

Maryna Popova[1], Larysa Globa[2] and Rina Novogrudska[1]
[1]National Center "Junior Academy of Sciences of Ukraine", Dehtyarivska Str. 38-44, 04119 Kyiv, Ukraine
[2]Igor Sikorsky Kyiv Polytechnic Institute, Peremohy avenue 37, 03056 Kyiv, Ukraine
pm@man.gov.ua, lgloba@its.kpi.ua, rinan@ukr.net

Keywords: Big Data, Multilevel Ontology, Taxonomization, Relations, Knowledge, Concepts.

Abstract: The problem of ever-increasing amounts of unstructured information in various fields of human activity is known as the problem of Big Data. Providing support for analytical activities requires determining the main factors that affect certain states of objects and processes in domains, as well as the degree of their influence, this significantly complicates the decision-making process, especially if data are represented heterogeneous information, there is a need to simultaneously take into account the impact of data from several areas dealing with several levels of classification. Given the significant volumes of text documents, it is impossible to solve the problem of structuring linguistic information by computer-aided extraction of the basic concepts that determine the text content (meaning), as well as the problem of constructing a formalized structure for formation the classes of individual objects and relations between them. The paper considers the ontological approach to the analysis and processing of Big Data represented both heterogeneous and linguistic data in the form of a multilevel ontology, implemented by computer-aided extracting of the basic concepts that define the text content (meaning) and determining semantic relations between the distributed information resources. The proposed approach uses the possibility of non-canonical conceptual ontologies to define equivalent concepts and thus to integrate the multiple ontologies that affect the same subject domain. This approach was implemented to create a multilevel ontology in the systemic biomedicine, the application of which in the process of postgraduate doctors and pharmacist's education has significantly reduced the search time of relevant information and errors number due to the lack of unified terminology.

1 INTRODUCTION

Today, in various fields of human activity, such as science, education, economics, health, business and other fields, there is so much data that the need to analyse and process them to improve the management of certain business processes is actual and urgently needed. It has stimulated the development of new intelligent data processing methods focused on practical application. An indisputable difficulty in solving various applied problems in different domains is the analysis and processing of Big Data that describe them and are characterized by diversity, large volumes, unstructured, as well as the inability to determine the degree of their impact on certain business processes, which, in turn, complicates decision-making processes. An even more complex problem is the decision-making process based on the processing of Big Data represented by linguistic information.

However, any human activity deals with domains that contain different components (sections), which are characterized by their own system of concepts, their knowledge and many tasks that require adequate formalized models of their representation. Ensuring the support of analytical activities requires the identification of the main factors that affect certain states of objects and processes in the domains, as well as the degree of their impact, so all available components relevant to specific tasks need to be integrated as some structure, such as a pyramidal network or graph. All these factors are important to represent and integrate at different levels. The need for simultaneous presentation of several domains that deal with several levels of classification, determined the development of the concept of multilevel modelling [1, 2].

In a broad sense, Big Data is a socio-economic phenomenon associated with the emergence of technological capabilities to analyse ultra-large data sets in some problem areas, but the entire world of a significant amount of linguistic (textual) information

requires significant automation of analytical processing [3], which consists in performing certain steps, namely: 1) structuring of linguistic (text) information due to computer-aided extraction (Data Mining, Data Extraction) of basic concepts that determine the content (meaning) of the text; 2) building a formalized structure by classes of concepts and relations between them formation; 3) determining the mechanism for conducting logical output based on the created structure.

The paper considers the ontological approach to the analysis and processing of Big Data represented by linguistic information in the form of a multilevel ontology, implemented by structuring texts and establishing semantic connections between distributed information resources.

The paper is structured as follows: Section 2 is dedicated to the analysis of the main data sources about knowledge representation in multilevel ontologies. Section 3 describes existed approaches to the implementation of ontologies, including multilevel ones. Section 4 provides an example of multilevel ontology implementation. The results of the paper are summarized in the Conclusion.

2 STATE OF ART

Today, ontological approach to automating the process of analytical processing of large amounts of textual information stored on the global network is wide spread [4-6].

Conventional ontologies (such as well-formalized OWL ontologies) use descriptive logic (first-order logic) to determine the class affiliation of individual objects and their binary relations (properties).

The concept of multilevel ontology or multidimensional ontology is to define a class of objects and their relations by geometric means in multidimensional space, not by formalisms based on first-order logic.

More formally, a multilevel ontology means a finite set of points that are classes in a multidimensional space. Multilevel ontology objects represented in the form of graph nodes are combined into classes according to certain properties, and binary relations («class-instance») – in the form of directed edges, connecting two objects. This representation of a multilevel ontology is present in almost all modern conceptual and ontological models. However, most of their use is limited due to the division of classes and instances into disparate sets. In other words, instances of classes cannot be other classes, and a multilevel taxonomy based on a class-instance relation cannot exist. Unlike first-order ontologies, binary relations between objects in a multilevel ontology are have no names. Similarly, they do not have semantically significant names and objects; instead, each is assigned a unique identifier (such as a URI in OWL ontologies).

Formally, any multilevel ontology can be reduced to an OWL ontology, but such a reduction will destroy the conceptual basis of the original multilevel ontology [7].

While creating intelligent systems based on knowledge, it is advisable to create a structure that would save the original form of the ontology, ensuring the integration of knowledge and ontologies of different domain sections. As a means of such integration, the authors suggest the use of meta-ontology, which defines the system of concepts described while creation of a domain sections ontology. Such a meta-ontology is an ontology of a more abstract level in relation to the domain sections ontology [8] (Figure 1).

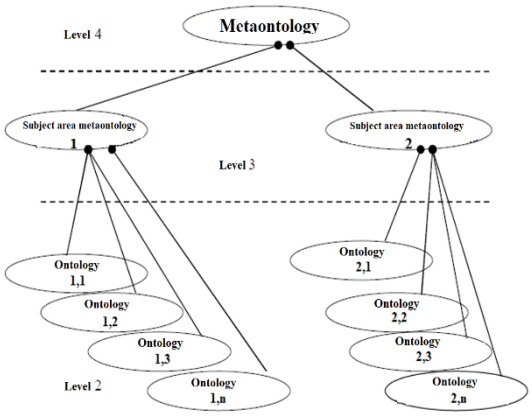


Figure 1: The multilevel ontology concept.

Metaontology, used to describe other ontologies, defines the structure of the internal organization of large ontology sections, indicating the general properties of the ontology matching types used in the domain section ontology. There are several levels of metaontology if this is necessary.

Multilevel integrated ontology (MIO) can be considered as an ontology, the concepts and roles of which are represented by dimensions, categories, measurements and facts. This ontology should also include all the axioms and statements needed to validate the intended model of multidimensional data. As a result, MIO can be used both to determine the directions of analysis and to test the resulting model for the presence of some new properties [9].

For knowledge management, multilevel ontologies and schemas for their representation are

considered mainly in the context of the data warehouses development and integration.

The authors [10] present a multilevel model with an OWL ontology model based on the descriptive logic of the Stanford Center for Biomedical Informatics Research, and define rules for transforming from a multidimensional level to an OWL ontology.

In [11] RDF-model OLAP-cube with an emphasis on the relation between the attributes of measures and measurements and its effect on the ability to summarize is described. The authors define the concept of measure- measurements consistency and demonstrate how to make logical output from OLAP ontology. The OLAP ontology is built using semantic web technologies and is mainly used to help users create OLAP cubes and queries to it.

Researchers from the Jaume I University (Spain) [9] propose to use ontology and semantically annotated data resources as a basis for designing semantic data repositories and an ontology-based environment for designing multidimensional analysis models.

In work [12] a new structure for the conceptual multilevel models' development, starting with a set of applied biomedical ontologies, is proposed. The methodology underlying the multilevel model is very simple, it is necessary to determine only the facts, indicators, measurements, categories and relations. This allowed to implement the model in almost any existing multidimensional database, performing the appropriate transformations. Regarding the scalability of the approach, the proposed solution allows you to manage large ontologies by selecting fragments that represent semantically complete modules of knowledge.

In the research [13] database transformation rules are used to generate the OWL ontology. Ontology-based technology provides semantic explanation and personalization capabilities based on the relation between concepts in the ontology. Multilevel ontology is designed to most fully reflect the terminology of a complex structured domain, identify common and partial in the content of such a complex structured area and provide the ability to reuse the description of concepts and relations in knowledge engineering and intelligent systems development.

The potential benefits of using multilevel ontologies are:

- obtaining additional levels of information presentation, which will be reused to create new levels;
- obtaining a more compact representation of the ontology text by introducing abstract concepts-relations between entities and their use in defining other terms.

Thus, the possibilities of multilevel ontology are used to solve the problem of semantic integration of reusable ontologies. In addition to the approach to finding relevant elements using metalevel specifications, the possibility of joining reusable ontologies as higher-level specifications can be considered. Acting as meta-information, such an ontology can remain independent and embedded without much efforts.

The multilevel ontology model should provide:

Facilitate the interpretation of concepts within the community. Today, there are many information systems characterized by the use of different conceptualizations, which complicates the interaction between them. Using a multilevel ontology will help solve this problem.

Reduction of errors. In many areas of human activity, such as medicine, there is no unified terminology, and the number of different ontologies is constantly increasing, which leads to semantic heterogeneity and, consequently, to the problem of semantic interoperability. However, the creation of a theoretical bridge in the form of a multilevel ontology will help to resolve ambiguities in ontological terms and concepts, thereby facilitating interpretation and reducing errors.

Data integration. Domain ontology is the only tool that allows you to reconcile at the semantic level of the model of heterogeneous data sources. Integration often occurs automatically because the ontologies used in the process capture and identify concepts in a formal and unique way.

Exchange of meaningful information. A coherent domain conceptualization can be easily used as a data exchange format. Unlike the usual exchange format, which defines the complete structure of the exchanged data and where the value of each data element is determined by its place in the global structure, ontology-based exchange is very flexible, which allows reasonable interpretation of completely different exchange structures by the same receiving system.

Extended support for semantic interoperability. Multilevel ontologies offer broader support for semantic interoperability, due to the fact that they reconcile the ontologies inconsistencies in different information systems.

Reuse of information. The ontology provides access to the data referenced by the concepts it defines. Ontologies are also used to query databases.

Existing approaches do not solve, first of all, the problem of structuring textual information by computer-aided extracting the basic concepts that determine the content (meaning) of the text, and the problem of building a formalized structure for forming classes of individual objects and relations between them. To solve these problems, it is proposed to automate the process of multilevel ontological model implementation by using software to solve analytical problems based on it.

3 MULTILEVEL ONTOLOGIES DESIGN

Ontology design is often not the ultimate goal in itself, usually ontologies are further used by intelligent systems to solve practical problems. In work [14] a 7-stage approach to ontology design is proposed.

1) Definition the scope and purpose of ontology.
2) Considering reusing existing available ontologies developed by someone else.
3) Listing of important ontology concepts, not taking into account possible coincidences of concepts that can be identified.
4) Definition of classes and their hierarchy.
5) Definition of properties related to classes.
6) Definition of constraints (number of elements, range of domain constraints), which relate to the properties.
7) Creating instances of classes in the hierarchy.

This approach uses the ability of non-canonical conceptual ontologies to define equivalent concepts and thus to integrate several ontologies describing the same domain.

An alternative approach to the ontology development starting with the canonical conceptual ontology is proposed in [15].

1) The first step in ontology development should be to agree in the community of its application. To reach an understanding, you should:
   - clearly define what is the domain described by the ontology;
   - choose a powerful model to accurately identify primitive concepts existing in the domain;
   - develop a common understanding of the canonical set of concepts that describe the field of knowledge.

2) Based on the certain canonical conceptual ontology, a non-canonical ontology can be created for practical use by a group of end users, to create their own idea of the domain or to formally model all concepts existing in the target domain related to ordinary linguistic notation (word or sequence of words). Thus, the possibility of exchanging information, expressed in concepts of the canonical conceptual ontology, is preserved.

3) To ensure that the ontology is used for linguistic output and/or to provide an end-user-friendly multi-lingual interface, it is necessary to define a list of concepts for a specific language and link them to each ontology concept. The multilevel "onion" model built based on this alternative approach [15] and obtained as a result of domain formalization, includes:
   - canonical conceptual ontology, which provides a formal basis for modelling (canonical and accurate descriptions of each concept) and effective exchange of knowledge in the domain between different sources;
   - non-canonical ontology, which provides mechanisms for linking various conceptualizations developed in this domain, which are used to interact with other software components or sources that already have their own special ontologies;
   - linguistic ontology, which represents the concept in natural language (in different languages) and sets the linguistic transformations over primitive and definite concepts.

Basic rules of ontology development according to [14] are formulated as follows:

1) There is no single right way to model the domain – there are always viable alternatives.
2) Ontology development is necessarily an iterative process – a repeated passage through the ontology in order to clarify it.
3) Ontology elements should be close to the objects (physical or logical) and relations in a particular domain.

Therefore, regardless of the choice of approach to multilevel ontology design, it is necessary to meet the basic requirements for its formation and development:

- flexibility – the ability to quickly and easily update any of ontology fragments, the ability to organize a decentralized "multi-agent" creation and editing of ontologies;
- openness – to add both individual concepts of any content and any conceptual subsystems, openness to the vocabulary of natural languages and additional options for

conceptual interpretation of words already contained in the lexicon of ontology;

- meaningful scalability – the ability to quickly select (expand/cut) certain fragments in accordance with the task, area of interest and point of view of individual professional groups;
- model scalability – the ability to present conceptual systems at different levels of detail to describe and formalize the relevant fragments of reality (for example, in the following sequence: simple semantic categorization of vocabulary – taxonomy – complete terminological model – production system – logical theory);
- versatility for the user – suitability for use in various software components and on different platforms.

4 MULTILEVEL ONTOLOGY IMPLEMENTATION

In this research the concept of ontology "level" is considered somewhat more widely.

First, the "internal" ontology level characterized by the depth of the binary relations of the taxonomy that underlie it is determined. The depth of binary relations means the depth of nesting of concepts categories, in graph terms it means that there is a certain distance between the terminal and root nodes, which exceeds 1 step: the greater the number of steps from root to terminal node, the higher the level of ontology.

Secondly, the concept of "external" ontology level characterized by the number of search iterations at the user request from the taxonomy concept context in the information sources integrated into the ontology is considered.

Multilevel ontology means a logical-linguistic model, the first level of which is represented by concepts in the form of logical formulas, reflecting the patterns inherent in the classes of objects and logical relations, the second – the corresponding concepts of consistent ontologies, the third and subsequent – semantically related information units contained in heterogeneous distributed sources of knowledge created by different standards and technologies and described in natural language (databases and knowledge bases, information banks, electronic archives, collections of electronic documents, etc.) (Figure 2).

4.1 Structuring Texts by Basic Concepts Computer-Aided Extraction

At the stage of first level of multilevel ontology formation the natural language texts taxonomization and contexts transdisciplinary categorization are carried out, that includes:

Concepts extraction – search in natural language documents terms that reflect the names, characteristics and relations between these terms.

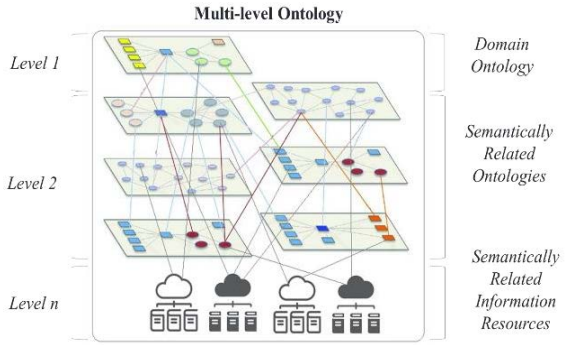


Figure 2: Multilevel ontology designing stages.

Usually, knowledge from any field of research is presented in text documents that contain poorly structured or even unstructured information. Processing such documents manually can be an extremely time-consuming process, and processing large arrays of such documents is almost impossible.

Before working with poorly structured or unstructured documents, it is necessary to structure them. During this process, the data is presented in a form convenient for computer-aided processing, which is easily read by standard means of ontologies designing, and is displayed in a user-friendly form.

The most difficult is to perform the structuring of natural language texts, as this process requires a sufficiently complete formal description of the language subset to which they belong. Each of the texts describes a specific area of research or part of it. The text uses the concepts that form its terminology. The text structuring consists of the concepts field extraction, in particular, the relevant field of study concepts (terms) identification, as well as their attributes and relations. The formed concepts field, in turn, can be represented as an ontology.

The task of text documents arrays structuring is to natural language process, that includes natural language semantic-linguistic analysis – the natural text documents processing, including formalization of the syntactic-semantic structure of sentences, computer-aided selection of multiword terms and

contexts in which they used, and given semantic relations based on templates of their descriptions.

The main result of semantic-linguistic analysis is the construction of a domain glossary – a list of objects that exist within it, that can act as either terms or names of certain entities. If the terms are read from the text, then in the future the text highlights the contexts in which these terms are used, and if possible – their definitions too. If named entities are read, then the process of determining the attributes of these entities is carried out in the future. Thus, the structuring of a certain natural language text T can be represented as a certain transformation (structuring transformation):

$$F_{str} : T \to O, \quad (1)$$

In fact, the text structuring transformation is a multi-stage process, each stage of which requires the use of specialized models and procedures. The process of text structuring can be divided into two main stages: lexical analysis $T \to T_{pr}$, which forms the primary structure of the text T_{pr}, and the ontology formation $T_{pr} \to O$, which allows you to select the necessary information from the primary structure and present it in the form of an ontology. To perform the second stage, a recursive reduction method is proposed [5, 16, 17], which provides a sequential primary structure transformation using a set of dynamically specified by the ontology designer rules.

The procedure of concepts extraction from natural language texts is implemented as one of the cognitive information technology "POLYHEDRON" [18] modules "KONSPEKT" [19], which functions include:

- text linguistic analysis to the level of superficial syntactic and semantic analysis;
- extraction of domain concepts from relevant texts;
- extraction and contextual description of the natural language texts concepts related to a given topic, which is given by a keyword or phrase;
- generation based on the results of semantic analysis of a given number of secondary keys, the use of which in a cyclic mode allows to deepen the disclosure of the topic in the formed contextual concepts descriptions;
- use contextual concepts descriptions to select from a set of text documents those that are most relevant to a given topic.

Terms from natural-language texts are distinguished using procedures and software tools for linguistic and semantic analysis of texts.

«KONSPEKT» [19] provides computer-aided contexts extraction that use the corresponding terms, and presentation of them in the form of a specialized XML structure. It allows computer-aided fill in the graph nodes with contexts based on the coincidence of nodes names and concepts names based on the results of semantic and linguistic texts analysis.

The results of semantic and linguistic analysis are used for natural-language texts taxonomization – a cognitive procedure for structuring text arrays based on the systemological representation of their terminological system in a hierarchical form. Because of natural-language text taxonomization, its structure can be represented as a graph, each node of which contains the corresponding contexts or attributes. Contexts content includes, respectively, semantic descriptions and characteristics of the corresponding concepts and phrases or characteristics of named entities.

Taxonomization provides extraction of classification units from text array that characterize its semantics and purpose. The text taxonomy reflects the order of interaction between terminological constructs or named entities.

Establishing relations between concepts. Relations indicate interactions between concepts. They are defined by properties and attributes that characterize domain classes.

Due to the established relations, ontology is not just a structure of concepts, but also reflects complex relations between them and comprehensively represents the domain. There are three main types of relations between concepts:

- R_t – taxonomic relations – express the relation «is-a» or the relation «general/partial»;
- R_c – compositional relations – express the relation «part of»;
- R_{top} – topological relations – reflect how different components of a terminological system are connected to each other through certain connections, or show the «paths» of physical interactions between components, as well as provide information about the spatial location of these components.

The definition of multiple relations of binary order over thematic concepts allows to achieve a high level of correctness in the formation of taxonomic categories and thematic classifiers. This ensures multiple interactions between taxonomic structures.

The result of applying the text taxonomization procedure is the definition of semantically significant relations between various objects, which can include

both certain relations between terms belonging to the domain (synonymy, class-subclass, etc.), and specific relations between named entities for this area of knowledge.

Representation of the primary text structure as a taxonomy. The taxonomy organizes concepts in a controlled dictionary into a hierarchy. The main purpose of the taxonomy is to create an ontology structure for human understanding and integration of other sources. In taxonomy, binary relations between different concepts of a domain are determined based on their definitions.

The primary text structure T contains a structured representation of lexemes (words or symbols), as well as syntactic relations between them. This structure, in fact, is an oriented graph, and lexemes are nodes of this graph.

Any natural-language text T is represented by a set of lexemes L, on which the precedence relations are defined $\Leftarrow$. This relation converts L to a linearly ordered set. The text T can also be represented as a sequence of sentences S that also define the precedence relation:

$$T = \left\{ S_1 \Leftarrow S_2 \Leftarrow ... \Leftarrow S_{n_i} \right\}, \quad (2)$$

where n_i is the total number of sentences in the text.

Each sentence S_i is represented by some subset of lexemes:

$$L_{S_i} = \left\{ l_{ij}, j = \overline{1, n_i} \right\}, \quad (3)$$

where n_i – the number of lexemes in *i* sentence.

Obviously, the condition is met:

$$\forall l_1 \in S_1, \forall l_2 \in S_2, S_1 \Leftarrow S_2 \Rightarrow l_1 \Leftarrow l_2, \quad (4)$$

where S_1, S_2 are arbitrary text sentences;

l_1, l_2 – lexemes.

Each lexeme has a number of attributes:

$$l_{ij} = \left\langle l_{ij}^T, P_{ij} \right\rangle, \quad (5)$$

where l_{ij}^T is the text representation of the lexeme l_{ij};

P_{ij} – lexeme attributes.

A lexeme can be related to other lexemes using syntactic relations $r_S \in R_S$:

$$r_S = \left\langle l^1, l^2, r_t \right\rangle, \quad (6)$$

where l^1, l^2 – lexemes that have a relation between them;

r_t – relation type.

Thus, oriented graph representing the primary structure of a natural-language text has the form:

$$T = \left\langle L, R_S \right\rangle. \quad (7)$$

The main problem is the inefficiency of working with the text representation of the lexeme, which is redundant and requires the construction of specialized functions defined on the set of words representations in text. Such functions are cumbersome and inefficient, and in software implementation, they often depend on the specifics of implementing text variable processing in a given programming language.

Since the set of text representations of lexemes is incalculable, we can construct a transformation of the form:

$$V: L^T \to \mathbb{N}, \quad (8)$$

where L^T is the set of text representations of lexemes.

Let the text be written in a specific alphabet A, the number of characters in which $n_{\mathrm{A}} = card(\mathrm{A})$. This alphabet can be considered as a notation with a base n_{A}. Accordingly, each letter $\alpha \in \mathrm{A}$ can be matched with a certain number $i_\alpha \in N$, which is the index of this letter in the alphabet. Any word in the input text is a sequence:

$$l^T = \left\{ \alpha_1, \alpha_2, ..., \alpha_{n_l} \right\}, \quad (9)$$

where n_l – word length $n_l > 0$;

α_i – letters of the alphabet A.

If we consider letters A as digits of a number in the corresponding notation, then such a number can be converted to decimal using the formula:

$$V(l^T) = i_{\alpha_1} \times (n_{\mathrm{A}})^{n_l} + i_{\alpha_2} \times (n_{\mathrm{A}})^{n_l - 1} + ... \\ ... + i_{\alpha_{nl}} \times (n_{\mathrm{A}})^0, \quad (10)$$

where i_{α_j} – letter index α_j in alphabet A;

n_{A} – number of characters in the alphabet A.

Using the V function, you can replace all l^T with their corresponding $l^V = V(l^T)$. As a result of this operation you can get a more efficient representation of the lexemes set:

$$\left\langle l^V, P \right\rangle \in L^V, \quad (11)$$

where l^V – code representation of a lexeme;

P – grammatical characteristics of the lexeme;

L^V – multiple code representations of lexemes.

In the future L^V, it can be considered as a set of lexemes L.

The taxonomy formation algorithm is based on the induction of utterances based on the selection of pairs (class name – name concept). If the statement is true, then a bipartite graph is constructed (a

unidirectional oriented graph with several edges entering and exiting one node). If the statement is false, then the graph is not constructed. The truth of the statement is established based on identifying the existence of a unifying property that is common to both concepts. The set of all bipartite graphs that are built on the set of true statements is a growing pyramidal network that is the basis of the taxonomy. The nodes contain class names and concept names.

Formally the technological basis for taxonomy formation is determined by a loaded bipartite graph:

$$G = (\hbar_1 \bigcup \hbar_2, E), \quad (12)$$

where $\hbar_1 \bigcup \hbar_2 = \varnothing$, nodes with $\hbar_1$ marked predicate names, and nodes with $\hbar_2$ marked argument names;

E – set of edges. Graph edges connect nodes marked with predicate names to vertices marked with argument names.

Vertices from the set $\hbar_1$ are called predicate nodes, vertices from the set $\hbar_2$ are called concept nodes, and predicates themselves are called conceptual predicates.

The statement is formed based on the composition of nodes incident to a single edge.

The ontological graph acts not only as a means of organizing information, but also as a environment for active user interaction with distributed information resources displayed in the form of a spatially ordered set of statements.

The effectiveness of using taxonomies in the process of integration and aggregation of information resources significantly depends on the quality of a domain structuring. Therefore, questions related to the ordering a set of taxonomic concepts determine the constructiveness of the knowledge system.

Axiomatization. Axioms provide the correct way to add Boolean expressions to ontology. Such logical expressions can be used to clarify concepts and relations in ontology. Axioms are used to develop an explicit way of expressing what is always true. Axioms can be used to determine the meaning of several components of ontology, identify complex relations, and verify the correctness of information or obtain new information.

Thus, the cognitive procedure for multi-stage sequential transformation of the primary text structure into an ontological form based on the selection of primary patterns – recursive reduction of natural language contexts that provides computer-aided transformation of text arrays into a taxonomy, thesaurus and ontology. The result of applying the procedure is the identification of lexemes (words or symbols, such as punctuation marks) that make up the attributes of domain objects (in particular, their names), the identification of primary intercontextual relations, and the taxonomic representation of text semantics.

The reduction process consists of sequentially extracting objects from the input text (a glossary of the domain is formed), relations between objects (domain taxonomy is formed) and attributes of objects (which are later considered as functions of interpretation (axioms), which allows us to consider the result of reduction as an ontology). This process can be represented by the following formula:

$$T \to O^1 \to O^2 \to O. \quad (13)$$

The recursive reduction method [16, 20] consists in recursively performing the process of reducing the input natural-language text, which, in turn, is carried out by applying a specialized operator to it:

$$F_{rd} : T \to O. \quad (14)$$

The reduction operator is a combination of four operators:

$$F_{rd} = F_{l*} \circ F_x \circ F_r \circ F_{ct}, \quad (15)$$

where F_{l*} is the aggregation operator that performs the auxiliary function of extracting phrases from the text that can represent a specific object; F_x is the operator for identifying ontology objects X. This operator applies a condition to the extracted phrases that determines whether to interpret a particular phrase as the name of an object; F_r – operator for identifying ontological relations R divided into relations between objects and auxiliary relations between the object and its contexts; F_{ct} – context identification operator that extracts its attributes from the context of a particular object (defined using the auxiliary relations extracted at the previous stage).

In general, each of the four transformation execution operators F is defined by the database of rules $DB_{\mathbb{R}}$ for performing this transformation. The rule RDBR has a unified structure for all stages:

$$\mathbb{R} = \langle f_{ap}^{\mathbb{R}}, f_{tr}^{\mathbb{R}} \rangle, \quad (16)$$

where $f_{ap}^{\mathbb{R}}$ – applicability function, which determines whether the rule can be applied to a specific set of input information;

$f_{tr}^{\mathbb{R}}$ – transformation function, which defines the transformation of input information.

The transformation $F_{\mathbb{R}} : X \to Y$ defined by the rule $\mathbb{R}$ has the form:

$$F_{\mathbb{R}}(x)=\begin{cases} f_{tr}^{\mathbb{R}}(x), f_{ap}^{\mathbb{R}}(x) \\ x, \neg f_{ap}^{\mathbb{R}}(x) \end{cases} \quad (17)$$

.

So, knowledge structuring by taxonomizing natural-language texts describing this knowledge to reflect the semantics of integrated and aggregated information resources in the form of hierarchical structures, over which a certain extensible axiomatic is defined and between which sets of relations are defined, allows us to solve the problem of their correct interpretation in the process of using ontology.

An important property of ontologies is the ability to structure information simultaneously with its perception. In this case, the formation of the memory structure occurs due to the interaction of perceived information and information that is already stored in the network graph. Because of the implementation of information structuring processes, the semantic and syntactic proximity of information is established. The found associative relations are fixed in the structural components of memory.

4.2 Ontological Interface Design

According to the ontological graph (taxonomy) model by means of computer-aided code generation by comparing the taxonomy objects with the set of source codes in the programming language was designed the ontological interface – a means of user-friendly interaction with the ontology [21, 22].

Changing the taxonomy (structure of the ontology) does not require making changes to the interface code, which provides dynamic extensibility, because it describes the correspondence between the ontology components and the target programming language instructions. Thus, the interface code generator is controlled by an ontology model, which is implemented as a wide set of software components and consists of static and dynamic parts. The static part contains file templates that implement fixed algorithms for controlling the code generation process, and the dynamic part contains algorithms for mapping descriptions of interface model components to program code (programming language instructions).

Ontological interface elements are the information content of a multilevel ontology. The visual representation of an ontology object is an image (drawing, picture, icon, photo, etc.), the source of which is specified in the corresponding node of the ontograph (taxonomy). The order of object display (in the form of an image gallery) of taxonomy concepts on the screen depends on the internal organization of nodes in the ontograph. The text description of the ontology object and links to sources of distributed information resources are displayed next to the image and have a common style for all objects (colour, size and font style, position in relation to the image, corresponding icons for links to information resources of various formats, etc.).

The ontological interface has tools for both horizontal and vertical navigation with elements of the slide menu and hamburger menu, which automatically adapts to different screen widths and mobile platforms. The “Prism” view mode uses full-screen navigation tools with the location of text and graphic elements of the ontology on 100% of the screen space. Therefore, ontological interface tools take advantage of the most common types of network resource navigation to reduce cognitive load and increase the efficiency of working with the ontology.

Based on the diversity of ontologies, establishing semantic agreement between them to ensure interoperability is a necessary condition for the formation of a second level multilevel ontology.

Ontology matching is the process of establishing a connection (conjunction) between different ontologies without changing the original ontology, so that both parties can get a common understanding of the same object [23]. It can also be defined as the process of finding a suitable object with the same or closest predictable value between two or more ontologies [24]. Otology matching takes two ontologies as input data and creates a semantic correspondence between entities in the two input ontologies. The authors [25] define ontological matching as follows: “Given two ontologies O_1 and O_2, matching one ontology with another means that for each entity (concept C, relation R, or instance I) in the O_1 ontology, we are trying to find a corresponding entity that has the same valid value in the O_2 ontology”.

Ontology matching can also be defined as a process in which two ontologies with overlapping content are linked at the conceptual level, and instances of the original ontology are computer-aided converted to instances of the target ontology according to existing relations [26].

Ontology matching is established after analysing the similarity of certain metrics of entities in comparable ontologies. The result of the ontology matching process is called alignment. Alignment is defined as a set of correspondences that represent relations between different entities. A match can be described by a tuple:

$$\langle id, s, e, r, v \rangle, \quad (18)$$

where id – unique match ID,

s – source ontology object O,

e – the essence of the target ontology O',

r – alignment relations such as equivalence (=), more general, intersection and disjunctionality of two entities and

v – reliability value, such as the similarity value.

Measurement of correspondence is the basis of all comparison algorithms, as it determines the degree of similarity between the ontologies to be matched. The formal designation of the similarity degree is given below [27]:

$$sim : E \times E \to R. \quad (19)$$

Similarity function:

$$E = E_1 \bigcup E_2, \quad (20)$$

where E_1 – a set of entities in the O_1 ontology,

E_2 – a set of entities in the O_2 ontology that receives two entities as input and calculates the similarity value.

The authors [28] provide some examples of similarity measurement that can be used when ontology matching. They include:

Terminological method that compares entity/concept labels. It uses purely syntactic approaches, as well as the use of vocabulary such as WordNet. The syntactic approach calculates correspondence using measurements of chain dissimilarity, while the lexical approach calculates correspondence using lexical relations such as synonymy and hyponymy.

Method for comparing internal structures that compares the internal structures of concepts, such as the value interval and attribute power.

Method for comparing external structures that compares relations between entities and other ontology components. It provides methods for comparing entities in an ontology and methods for comparing external structures by taking into account loops.

Semantic method compares interpretations or entities models in ontology.

To form the third and subsequent levels of a multilevel ontology, indexing of a set of information resources (Big Data sources) was performed using lexicographic systems virtualization technology [29] and an agent approach.

Thus, the third and subsequent levels of multilevel ontology are digital collection of documents formed as a result of systematization of network resources, a set of natural-language texts united by one or a set of features (linguistic, conceptual, pragmatic, temporal, stylistic, functional, intentional, etc.). The most popular are collections of texts with the same topic (educational and scientific collections), one author (complete works), a certain historical era, a certain language or created under certain circumstances, in a certain form, for a certain purpose (educational and methodological materials, normative legal acts regulating legal relations in a certain area, etc.), for a category of readers with a certain level of access (public data, data for official use), etc. Modern information and communication technologies allow doing this dynamically, selecting full-text documents relevant to the user's request from a Web supermassive of indexed texts or local databases – specialized electronic libraries.

An important tool for studying text collections is the relation of semantic identity of natural-language texts.

While forming online digital collections of text documents, the following methods are used: comparative analysis (checking texts for semantic identity), system analysis (researching semantic identity as a relation with system-forming properties) and modelling the relation of semantic identity of natural-language texts.

The natural-language text expertise is based on the representation of a natural-language text as an “ordered hierarchy of content objects”.

Theory of lexicographic systems [30] operates with the concept of elementary information units (EIU), which it interprets as a subsystem of relatively stable discrete entities that is induced in the structure of any system and develops as a result of the action of various types of L-effects. Accordingly, all non-elemental objects of the system are considered as certain combinations of EIU.

During processing, a two-level hierarchy of text content objects is set. At the first (“upper”) level fragments of text, “thought blocks” that reveal one topic (micro theme), at the second (“lower”) – their constructive units (components) – thought objects, carriers of subject meanings and relations – words, phrases, combinations of words. Thus, the components of fragments act as elementary information units.

Concepts that denote the same subject are identical. At the same time, if the volume and the same generic feature completely coincide, they have different content and differ in species characteristics.

Creating collections of network texts in a multilevel ontology is provided by using specialized technologies of ontology-driven web or intranet crawling. The crawler subsystem is tightly integrated with the corps system and indexing system and the

multi-language synonym zone. Crawlers, like the corps system, are virtualized lexicographic agents, they are a type of lexicographic systems.

The identity search methodology involves pre-processing texts to improve the efficiency and speed of search – normalization.

Searching for text identities on the Internet using web crawlers is a non-trivial task. The search scenario is as follows:

- the length of queries is determined (8 words in a series are considered optimal);
- queries are generated from eight overlapping words (1 – 8; 2 – 9; 3 – 10 etc.). A single-word overlap gives you the highest search quality, although it takes longer to complete the query, so for large texts, the overlap should be as small as possible. Requests refer to the crawler API.

The array of texts received in response to the query contains a significant amount of search noise – mistakenly defined as text identities, so it is subject to further processing.

In further processing, suffix trees, the thesaurus method with a multi-language synonymous zone, the shingles method, Bag of Words, the N-Gram method, and distributive semantics are used.

Described methodology usage was tested while developing multilevel ontology "Systemic biomedicine". Its structure is shown on Figure 3.

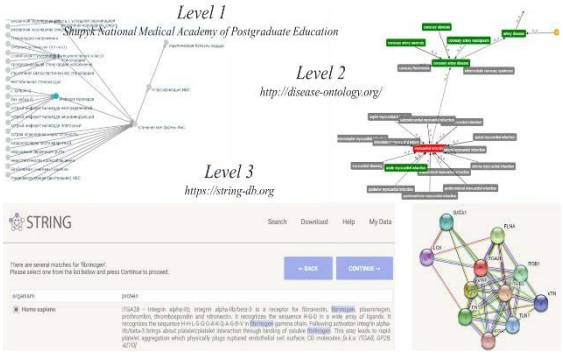


Figure 3: Example of implementing the multilevel ontology "Systemic biomedicine" scheme.

Each of the levels of a multilevel ontology can be flexibly expanded and supplemented with new objects, as well as integrate distributed information systems and sources of information resources.

In practice, multilevel information classification is not introduced in many data models. The main reason lies in the high computational complexity of logical problems associated with its modelling. The description of a multilevel classification cannot be modelled in first-order logic, since the metaclass is modelled by a second-order statement in which the variable bound by the quantifier must take values corresponding to the classes.

Thus, the ontology description language Ontolingua [31] is based on first-order logic, given the use of the KIF language for statements, and therefore does not allow multilevelness from the very beginning. The ontology language in Semantic Web Technologies OWL uses the RDF Schema language as its basis, which allows modelling of metaclasses. The OWL Full dialect uses this feature, but it is not allowed. The OWL DL dialect does not preserve the semantics of RDF Schema classes, but introduces its own, corresponding to descriptive logic, which assumes a subset of first-order constructs in which you can solve feasibility problems and some other logical problems. The same applies to OWL 2 language profiles and their corresponding logics. None of the profiles introduces the possibility of modelling metaclasses, despite the fact that approaches are used to increase the expressive power, in particular, in the RL profile – by introducing conditions for the use of constructs in statements of superclasses and subclasses [32]. Therefore, to create a multilevel ontology, a specific XML format is used that can ensure the interoperability of information at all levels to implement multilevel ontology.

5 CONCLUSIONS

Carrying out research has shown that there are no approaches to solve the problem of structuring text information by computer-aided extraction of the basic concepts that determine the text, content (meaning) as well as the problem of constructing a formalized structure for formation the classes of individual objects and relations between them.

The solution to the problem of significant linguistic information amounts analytical processing available on the Internet is proposed.

It consists of performing such steps of texts processing as structuring texts by computer-aided extraction of basic concepts that determine the content (meaning) of the text; building a formalized structure for the individual objects' classes and relations between them formation; determining the mechanism for conducting logical output based on a multilevel ontology.

The modified model of a multilevel ontology differs from the known ones in that the concept of a level is considered not in three-dimensional space, but in multi-dimensional space due to the semantic connectivity of distributed information resources.

The multilevel ontology model is able to provide facilitating the interpretation of concepts within the community, reducing errors due to the lack of unified terminology, data integration, exchange of meaningful information, extended support for semantic interoperability, information reuse.

In the future, it is planned to use the proposed multilevel ontology model to improve the mechanisms for searching and conducting logical inference.

REFERENCES

[1] B. Neumayr, K. Grn, and M. Schrefl, "Multilevel domain modeling with m-objects and m-relationships," Proc. of 6th Asia-Pacific Conf. on Conc. Model., New Zealand, 2009.

[2] C. Atkinson and T. Khne, "The essence of multilevel modeling," Proc. of 4th Int. Conf. on the Unified Model. Lang., pp. 19-33, Toronto, Canada, 2001.

[3] V. Mayer-Schonberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Canada: Eamon Dolan/Houghton Mifflin Harcourt, 2013, p. 242.

[4] A. Luntovskyy and L. Globa, "Big Data: Sources and Best Practices for Analytics," Proc. of International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo'19), pp. 1-6, 2019.

[5] O. Stryzhak, V. Prychodniuk, and V. Podlipaiev, "Model of Transdisciplinary Representation of GEOspatial Information," in Advances in Information and Communication Technologies. Lecture Notes in Electrical Engineering, vol. 560, Cham: Springer, 2019, pp. 34-75

[6] M. Popova, O. Stryzhak, O. Mintser, and R. Novogrudska, "Medical Transdisciplinary Cluster Development for Multivariable COVID-19 Epidemiological Situation Modeling," Proc. Of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2020), 2020, pp. 1662-1667, doi: 10.1109/BIBM49941.2020.9313204.

[7] G. Barzdins, N. Gruzitis, G. Nešpore-Bērzkalne, B. Saulīte, I. Auziņa, and K. Levāne-Petrova, "Multidimensional Ontologies: Integration of Frame Semantics and Ontological Semantics," Proc. of 13th EURALEX International Congress, pp. 23-28, Barcelona, Spain, 2008

[8] I. L. Artemyeva, "Complexly structured subject areas. Construction of multilevel ontologies," Information Technology, vol. 1, pp. 16-21, 2009.

[9] V. Nebot, R. Berlanga-Llavori, J. Pérez-Martínez, M. Aramburu, and T. Pedersen, "Multidimensional Integrated Ontologies: A Framework for Designing Semantic Data Warehouses," Data Semantics, vol. 13, pp. 1-36, 2009.

[10] N. Prat, J. Akoka, and I. Comyn-Wattiau, "Transforming multidimensional models into OWL-DL ontologies," Proc. of Multidimensional Models Meet the Semantic Web: Defining and Reasoning on OWL-DL Ontologies for OLAP, Hawaii, USA, 2012.

[11] T. Niemi and M. Niinimäki, "Ontologies and summarizability in OLAP," Proc. of the Proceedings of the 2010 ACM Symposium on Applied Computing(SAC'10), pp. 1349-1353, March 2010.

[12] O. P. Mintser and V. M. Zaliskiy, Systemic biomedicine. Kyiv: Interservice, 2019, p. 552 .

[13] L. El Saraj, B. Espinasse, T. Libourel, and S. Rodier, "Towards Ontology-Driven Approach for Data Warehouse Analysis," Proc of 8th Int. Conf. on Software Eng. Advances (ICSEA 2013), pp. 1-6, Venice, Italy, 2013.

[14] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Technical report ksl-01-05 and stanford medical informatics technical report smi-2001-0880, Stanford Knowledge Systems Laboratory, 2001

[15] S. Jean, G. Pierra, and Y. Ait-Ameur, "Domain Ontologies : A Database-Oriented Analysis," Proc. of Web Inf. Sys. and Techn. (WEBIST'2006), pp. 238-254, Set`ubal, Portugal, April 2006.

[16] V. Prychodniuk, "Technological means of transdisciplinary representation of geospatial information," ITGIS, Kyiv, 2017.

[17] V. Prychodniuk, "Taxonomy of natural-language texts," Information Models and Analyses, vol. 5(3), pp. 270-284, 2016.

[18] O. Ye. Stryzhak, L. S. Globa, V. Y. Velichko, M. A. Popova, and others, "Computer program Cognitive IT platform "POLYHEDRON"," Certificate of copyright to the work №96078 dated 17.02.2020, Official bulletin No 57 (31.03.2020), pp. 402-403, 2020.

[19] V. Velychko, M. Popova, V. Prykhodniuk, and O. Stryzhak, "TODOS – IT-platform for the formation of transdisciplinary informational environments," Syst. Arms and Milit. Equip., vol. 1(49), pp. 10-19, 2017.

[20] O. Ye. Stryzhak, V. V. Prykhodniuk, S. I. Haiko, and V. B. Shapovalov, "Display of network information in the form of interactive documents. Transdisciplinary approach," Math. Model. in econ., vol. 5(3), pp. 87-100, 2018.

[21] M. Popova, "Ontology of interaction in the environment of the geographic information system," ITGIS, Kyiv, 2014.

[22] M. Popova, "A model of the ontological interface of aggregation of information resources and means of GIS," Inf. Tech. and Knowl., vol. 7(4), pp. 362-370, 2013.

[23] C. Rung-Ching, L. Bo-Ying, and B. Cho-Tscan, "Using Domain Ontology Mapping for Drugs Recommendation," Department Of Information Management, Chaoyang University Of Technology, Taiwan, 2009

[24] I. Olaronke, A. Soriyan, and I. Gambo, "Ontology Matching: An Ultimate Solution for Semantic Interoperability in Healthcare," Int. J. Comp. App., vol. 51, pp. 7-14, 2012, doi:10.5120/8325-1707.

[25] M. Ehrig and S. Staab, "QOM – Quick Ontology Mapping," Proc of the Int Sem. Web Conf., vol. 3298, pp. 683–697, 2004.

[26] B. Veli, B. L. Gokce, D. Asuman, and K. Yildiray, "Artemis Message Exchange Framework: Semantic Interoperability of Exchanged Messages in the Healthcare Domain," Software Research and

Development Center, Middle East Technical University (METU), Ankara, Turkiye, 2006.

[27] S. Z. Katrin, "Instance-Based Ontology Matching and the Evaluation of Matching Systems," Inaugural-Dissertation. Department of Computer Science, Heinrich Heine University of Dusseldorf, Germany, 2010

[28] E. Rahm and P. Bernstein, "A Survey of Approaches to Automatic Schema Matching," VLDB Journal, pp. 334-350, 2001.

[29] M. V. Nadutenko , "Virtualized lexicographic systems and their application in applied linguistics," ULIF, Kiev, 2016.

[30] V. A. Shirokov, Information theory of lexicographic systems. Kyiv: Dovira, 1998, p. 331.

[31] The ontology description language Ontolingua [Online]. Available: http://www.ksl.stanford.edu/ software/ontolingua, July 2005.

[32] A. E. Vovchenko, V. N. Zakharov, L. A. Kalinichenko, D. Yu. Kovalev, O. V. Ryabukhin, and oth., "Multilevel specifications in conceptual and ontological modeling," Proc. of 13th All-Russian Scientific Conf. (RCDL'2011), pp. 35-43, Voronezh, 2011.

Influence of Synthetic Image Datasets on the Result of Neural Networks for Object Detection

Aleksandr Kniazev[1, 2], Pavel Slivnitsin[1, 2], Leonid Mylnikov[1], Stefan Schlechtweg[2] and Andrey Kokoulin[1]
[1]*Perm National Research Polytechnic University, Komsomolsky avenue 29, 614990 Perm, Russian Federation*
[2]*Anhalt University of Applied Sciences, , Bernburger Str. 57, 06366 Köthen, Germany*
knxandr@rambler.ru, slivnitsin.pavel@gmail.com, leonid.mylnikov@pstu.ru, stefan.schlechtweg@hs-anhalt.de, a.n.kokoulin@gmail.com

Keywords: Image Recognition, Object Detection, Neural Network, Synthetic Dataset, Data Generation.

Abstract: The goal of the article is research of ways to improve the quality of neural networks object detection. To achieve this goal we suggest to use synthetic image datasets. The algorithm of generating synthetic images, which uses the environment of the detected object, is described in the article. That algorithm could be applied in the control algorithm of the robotic system for luminaire replacement that is based on target object detection. 3D models and 3D camera images of detected objects, backgrounds, noise objects and different effects are used to create realistic images that will increase the quality of predictions. Quality tests were made with synthetic and real datasets. Results show that quality could be increased up to 16%. Ratio of real and synthetic data is 1:4.

1 INTRODUCTION

Training is a very important part of neural network creation. Less datasets leads to undertraining, while huge datasets leads to overtraining. Even optimal size of dataset can lead to bad results if objects for detection would be captured from one view or/and on the same background. Moreover false positive detection can appear. The order of the training dataset is also an important thing in the training process [1]. In case of object detection images annotated with coordinates of objects are elements of training datasets.

Manual annotation is a very popular way to annotate images presently. Scientists have to define bounding boxes of objects by hand in special programs (for example LabelImg - https://github.com/tzutalin/labelImg). Example of manual annotation is shown on Figure 1. This process is very time consuming.

Community of scientists created a huge amount of annotated datasets for the last 20 years [2]. These datasets can be easily found and have free access to use in tasks of object detection. However, all these datasets are applicable for a small range of object detection tasks and cannot be applied in other tasks. Luminaire detection is one of these tasks which demand dataset creation. Neural network for luminaire detection can be applied to a robot which replaces broken luminaires [3] with a special connector [4].

Figure 1: Manually annotated image with LabelImg.

There are some articles about automatic generation of training datasets in literature. Some simple methods use a sliding window to capture movable objects [5]. Other methods use combinations of elements to generate images. For example, 3D models of objects for detection are used with different backgrounds [6-8]. Moreover, additional noise effects can be applied to simulate different factors which can influence image quality [9]. This can improve precision of detection.

To generate datasets in [6-7] each 3D scene has to be manually set in Blender 3D. Complex algorithms of calculating horizontal planes are used in [10]. After that, scientists have to manually remove false regions. Finally, an image would be generated. All these factors strongly influence the speed of generation.

Synthetic datasets can influence the quality of neural networks results. To measure this influence it is necessary to use methods of assessing the accuracy of object detection. Intersection over Union (IoU) also known as Jaccard index [11] is a widespread method. This method compares two shapes. That is why IoU is invariant to the scale of the object in the

image. Due to this property the precision of the detected object is measured [12, 13].

2 METHODOLOGY

To generate a training dataset we will use a combination of background image, image of the detected object and some noise objects. A small amount of images could be enough to create big datasets. Random position, size of detected object, different position, gamma and blur value of noise objects allow to create a lot of various datasets.

To obtain an image of the detected object we use two approaches: 3D model and depth image from a 3D camera. Figure 2 shows an example of the algorithm of generation synthetic images.

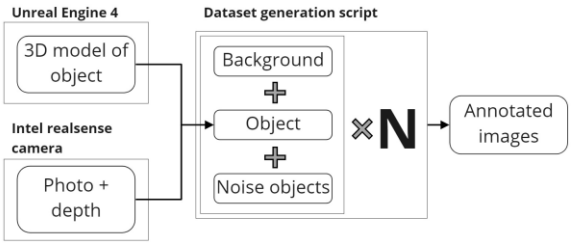


Figure 2: Dataset generation algorithm (N - is a number of images).

The advantage of this approach is that we know positions of detected objects and can automatically annotate images. Blender 3D is used to obtain images of the detected object. Firstly, a 3D model has to be created. Then the 3D model is rotated by Z-axes and rotation is recorded as animation. Animation then should be saved as a set of images. To calculate object position background should have a color which is contrast to the detected object.

Figure 3: Examples of images of objects obtained from 3D models.

Another approach is to use a 3D camera to obtain images of detected objects. We used an Intel RealSense camera which can capture simple RGB images and in addition it captures depth of images. Depth image is used as mask to separate object and background. Then mask have to be written to the alpha channel of the PNG image. It allows to combine object image and background.

Such images could have some defects due to transparent parts of objects. For example, a light bulb of the luminaire. Figure 4.a, shows these defects.

To fix these defects we use the Graph Cut method [13]. Example is shown on Figure 4.b.

a) b)

Figure 4: Examples of object images obtained with 3D camera: a) image of object directly from 3D camera, b) image of object after Graph Cut method.

Generated images of objects should be placed on background images in a realistic way. It means that the background should correspond to places in the real world where objects could be placed. Noise objects then placed in addition. In case of luminaires, it could be tree branches, which can cover luminaire, rain, low light, camera defects. These noise effects make the resulting image more realistic. To make a more realistic result we add some blur to noise images.

Mask is calculated to combine object image and background. Coordinates of the object set randomly from predefined parameters. After that, the object image is put on the result image. Moreover, gamma and size changes can be applied. Noise objects are preprocessed the same way. In addition, noise objects can be flipped.

Object mask is also used to annotate images. As a result "xml" file is obtained (example is shown in Listing 1) with coordinates of object bounding box. This box is used in neural network training.

Listing 1. Annotation of an object on the image in the form of an "xml" file.

```
<annotation>
  <folder>train</folder>
  <filename>1608936293.0034409.jpg</filename>
  <path>test_saves/1608936293.0034409.jpg</path>
  <source>
    <database>LumAutoGenDataset</database>
```

```
  </source>
  <size>
    <width>416</width>
    <height>416</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>luminaire</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>250</xmin>
      <ymin>117</ymin>
      <xmax>320</xmax>
      <ymax>264</ymax>
    </bndbox>
  </object>
</annotation>
```

This approach allows to obtain a huge amount of unique images. Examples of generated images are shown on Figure 5.

Figure 5: Examples of images from generated datasets.

3 EXPERIMENTS

YOLOv3-Tiny [14] was chosen to test the influence of synthetic data on neural network results. This neural network model demands less time on training and testing. It allows to make more experiments and compare results. Model structure is simpler than other models [15]. We can suggest that the influence of synthetic datasets will show and it could appear on other models.

We used mixed datasets of real photos and synthetic images.

Generation of synthetic datasets was made by the algorithm on Figure 6.

Step 1: Setting the required number of generated images (N);
Step 2: Loading of N random backgrounds from predefined set of 2D images;
Step 3: Loading of object images from predefined set of images with 3D models and images from 3D camera for each background;
Step 4: Random changes of gamma, size and position of objects;
Step 5: Add object images on backgrounds by mask;
Step 6: Calculation of object bounding box and saving to "xml" file;
Step 7: Loading of noise objects k times for each background ($k \in (0; m)$, m – is the maximum number of noise objects on en image). Calculation of masks of noise images;
Step 8: Random change of gamma, size and position of each k noise object for each background;
Step 9: Add k noise objects to backgrounds by masks;
Step 10: Save all generated images.

Figure 6: Algorithm of generation synthetic datasets for training neural networks (to work with random numbers we use uniform distribution law).

Since while preparing the training data we do not have information about the location of the target object and do not take it into account, we will assume that if the data is mixed evenly, we can avoid a drastic change in weights during neural network training.

The learning quality in this case depends on how evenly our data is shuffled and how diverse the data is generated. When these operations are performed optimally, the result can be expected to be non-random according to the central limit theorem [16].

After synthetic data generation, the parameters that affect the quality of the model are: 1) ratio of synthetic data and real photos in the training dataset; 2) size of training dataset.

This results in two criteria:

1) quality criterion $\mathrm{IoU}(\Omega(x, y))) \rightarrow \max$,

2) time criterion, $T(\Omega(x, y))) \rightarrow \min$, where Ω is training dataset, x is the number of real data, y – is the number of synthetic data, T – training time.

The criteria are differently oriented. As a result of empirical research (using different data ratios of 1:4, 1:8 and dataset sizes of 1000 and 2000 images), a real to synthetic data ratio of 1:4 and a dataset size of 1000 images were chosen.

For the completeness of the research, we tried to train the model using various combinations of real and synthetic data: training using synthetic data with 3D model images and validation on similar data; training using synthetic data with 3D model images

and validation on real images; training using mixed data (3D models + real photos) and validation on real images; training using 3D camera images and validation on similar data; training using 3D camera images and validation with real data.

Table 1: Synthetic datasets for training.

№	Number of real photos	Number of generated images
1	200 + 50 (training, validation)	0
2	0	800 + 200 (training, validation) (3D model)
3	250 (validation)	1000 (training) (3D model)
4	100 + 150 (training, validation)	1000 (training) (3D model)
5	100 + 150 (training, validation)	1800 + 200 (training, validation) (3D model)
6	250 (validation)	1000 (training)
7	100 + 150 (training, validation)	1000 (training) (3D camera)
8	250 (validation)	1000 (training) (3D camera + Graph Cut)
9	100 + 150 (training, validation)	1000 (training) (3D camera + Graph Cut)
10	0	800 + 200 (training, validation) (3D camera + Graph Cut)
11	0	2400 + 600 (training, validation) (3D model, 3D camera, 3D camera + Graph Cut)

All models were trained in the same way, only the input data was changed. Model YOLOv3-Tiny pre-trained on the COCO trainval dataset provided by the developers on the official website [17] was used for experiments.

The training consisted of two stages, the first one with frozen weights of all layers except the last two ones, responsible for object detection. This was performed to obtain stable losses to reduce the impact of the initial high losses on the weights in the main part of the model. After the losses stabilisation within 50 epochs, all weights were unfrozen and training continued.

After first attempts to train the neural network on synthetic data obtained using 3D model, we have observed that training and validation of the model using only synthetic data gives poor quality of object detection in real photos, an example is shown in Figure 7.

For this reason, we decided to also train models on synthetic data with validation on real data and on mixed data.

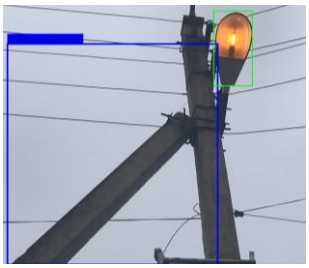


Figure 7: An example of poor object detection with a model trained on only synthetic data.

4 RESULTS AND DISCUSSION

After training the models with the datasets described in Table 1 and evaluating their performance on IoU metric (see Tables 2 and 3), we can say that there is a 0.002-0.16 improvement in neural network performance for 5 experiments and 0.013-0.146 for 10 experiments using datasets 5, 6, 7, 9 compared to 1. The other datasets had no positive effect on the performance of object detection in the image. Increasing the number of experiments, there is a slight fluctuation in the dispersion, which indicates reproducibility and result stabilisation despite the stochastic elements used in the models according to the central limit theorem.

The lowest quality is shown by training the model using 3D models with validation on real photos. It can be caused by the fact that the 3D models we use have technical inaccuracies (models may not look enough detailed and believable from some angles). The validation data were realistic and detailed, that may have reduced the detection quality of the model.

Figures 8 to 10 show the loss curves for models (2), (3), (9) compared with model (1).

The curves show the difference in the effect of validation data on learning. For example, model (3), which was trained using real photos for validation, significantly reduced its performance after unfreezing the main part of the YOLOv3-Tiny neural network weights. Whereas the learning curve of model (2), which was validated on the same type of data as the

training data, did not show such a sharp increase in losses. This allows us to see that a network trained on 3D model images has learned to detect luminaires in the generated images (3D model), but it will not be able to detect luminaires in real photos.

Table 2: Average IoU value and dispersion of IoU values for 5 experiments.

№	Average IoU value	Dispersion of IoU values
1	0,427	0,136
2	0,081	0,042
3	0,217	0,076
4	0,429	0,142
5	0,429	0,146
6	**0,488**	0,141
7	**0,587**	**0,098**
8	**0,431**	**0,112**
9	**0,592**	**0,091**
10	0,413	0,11
11	0,429	0,137

Table 3: Average IoU value and dispersion of IoU values for 10 experiments.

№	Average IoU value	Dispersion of IoU values
1	0,444	0,131
2	0,059	0,032
3	0,249	0,086
4	0,413	0,145
5	**0,457**	0,148
6	**0,477**	0,146
7	**0,59**	**0,101**
8	0,419	0,111
9	**0,586**	**0,091**
10	0,428	0,113
11	0,434	0,14

The lack of a sharp increase in losses after weights unfreezing can be observed in Figure 10 (model 9). This is due to the fact, that the validation was performed using both synthetic images and real photos. This improved quality of object detection in cases (7) and (9).

Neural network training based on the generated dataset using Intel RealSense 3D camera images showed better results compared to the 3D model images. The combination of synthetic images and a small number of real photos improved the quality of object detection compared to models trained on only a small number of real photos. This shows that this approach can be used to improve the quality of object detection.

The research described in the paper focused on the detection tasks of static objects. However, we assume that since moving objects can be represented by a set of sequential images, the proposed approach can be extended to moving images as well. This is possible by using an algorithm that makes the necessary corrections for false positives or false negatives of the base algorithm on single images.

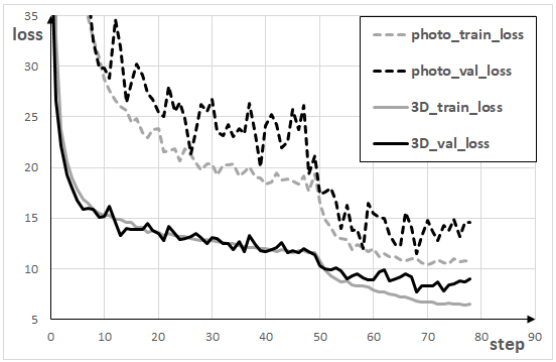


Figure 8: Loss curves for the model trained on synthetic dataset 2 (see Table 1).

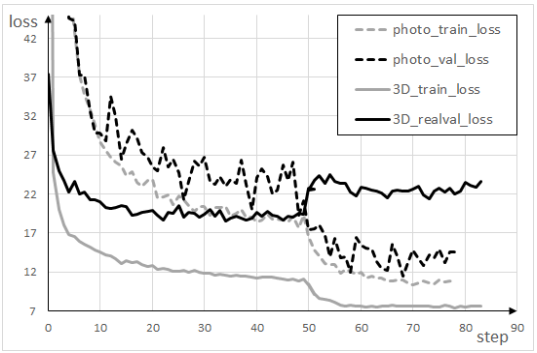


Figure 9: Loss curves for the model trained on synthetic dataset 3 (see Table 1).

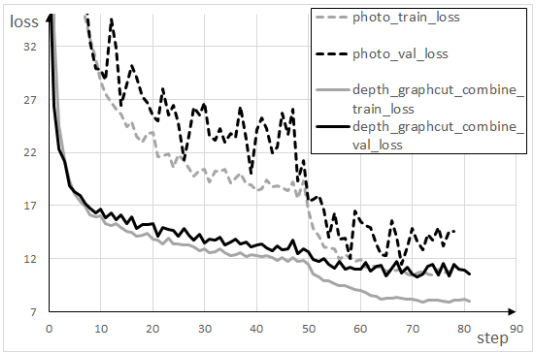


Figure 10: Loss curves for the model trained on synthetic dataset 9 (see Table 1).

5 CONCLUSION

Data sets of 1,000 images each were used in the experiments. As a result, we have found that it is possible to create datasets on synthetic data, but it is also necessary to dilute this synthetic image dataset with a small number of real photos (with a ratio of real to synthetic data approximately ¼). This solves the problem of creating large annotated datasets, required for training neural networks to improve the quality of object detection. This paper shows the effect of different combinations of synthetic and real data on the performance of a neural network for object detection. In our paper, we have tried to perform as many experiments as possible to get the broadest possible overview of the impact of synthetic data on neural network performance.

The proposed approach differs from existing approaches by using a combination of 3D models, fragments of real photographs and noise effects. In addition, this approach does not use algorithms that calculate the position of objects in 3D space and algorithms that calculate the possible position of detection objects. In our example (luminaire detection), the object can be located in any part of the image. This reduces the required time to create a single image for a dataset.

We expect that works intended to produce more realistic images, for example containing elements such as corrosion and deformation effects, and failures will contribute to further improvements in detection quality.

REFERENCES

[1] L. Mylnikov, “Statistical methods of intelligent data analysis,” St. Petersburg: BHV-Petersburg, 2021, 240 p.

[2] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object Detection in 20 Years: A Survey,” arXiv, pp. 1-39, May 2019.

[3] P. Slivnitsin, A. Bachurin, and L. Mylnikov, “Robotic system position control algorithm based on target object recognition, ” in Proceedings of International Conference on Applied Innovation in IT, vol. 8, no. 1, pp. 87-94, 2020.

[4] P. A. Slivnitsin and A. A. Bachurin, “A modern way of outdoor lighting maintenance, ” in Journal of Physics: Conference Series, vol. 1415, no. 1, 2019.

[5] T. Anwar, “Training a Custom Object Detector with DLIB & Making Gesture Controlled Applications,” 2020 [Online]. Available: https://www.learnopencv.com/training-a-custom-object-detector-with-dlib-making-gesture-controlled-applications/ [Accessed: 07-Dec-2020].

[6] J. Li, P. L. Götvall, J. Provost, and K. Åkesson, “Training Convolutional Neural Networks with Synthesized Data for Object Recognition in Industrial Manufacturing,” IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA, vol. 2019-Septe, pp. 1544-1547, 2019.

[7] M. Andulkar, J. Hodapp, T. Reichling, M. Reichenbach, and U. Berger, “Training CNNs from Synthetic Data for Part Handling in Industrial Environments,” IEEE Int. Conf. Autom. Sci. Eng., vol. 2018-Augus, pp. 624–629, 2018.

[8] D. Mas Montserrat, Q. Lin, J. P. Allebach, and E. J. Delp, “Scalable Logo Detection and Recognition with Minimal Labeling,” Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018, pp. 152-157, 2018.

[9] G. Volk, S. Muller, A. Von Bernuth, D. Hospach, and O. Bringmann, “Towards Robust CNN-based Object Detection through Augmentation with Synthetic Rain Variations,” 2019 IEEE Intell. Transp. Syst. Conf. ITSC 2019, pp. 285-292, 2019.

[10] G. Georgakis, A. Mousavian, A. C. Berg, and J. Košecká, “Synthesizing training data for object detection in indoor scenes,” Robot. Sci. Syst., vol. 13, 2017.

[11] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June, pp. 658-666.

[12] G. T. U. A. Colleges, et al., “Microsoft COCO,” Eccv, no. June, pp. 740-755, 2014.

[13 M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303-338, 2010.

[14] J. Redmon and A. Farhadi, “YOLO v.3,” Tech Rep., pp. 1-6, 2018.

[15] W. He, Z. Huang, Z. Wei, C. Li, and B. Guo, “TF-YOLO: An improved incremental network for real-time object detection,” Appl. Sci., vol. 9, no. 16, 2019.

[16] P. G. Doyle, “Grinstead and Snell's Introduction to Probability,” 2006, American Mathematical Society. 518 p

[17] J. Redmon and A. Farhadi, “YOLO: Real-Time Object Detection,” 2018 [Online]. Available: https://pjreddie.com/darknet/yolo/. [Accessed: 25-Nov-2020].

Urban Environment Simulator for Train Data Generation Toward CV Object Recognition

Kirill Karpov[1,2], Ivan Luzianin[1], Maksim Iushchenko[1,2] and Eduard Siemens[1]
[1]*Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*
[2]*Department of Transmission of Discrete Data and Metrology, Siberian State University of Telecommunications and Information Sciences, Kirova Str. 86, 630102 Novosibirsk, Russian Federation*
{kirill.karpov, ivan.luzianin, maksim.iushchenko, eduard.siemens}@hs-anhalt.de

Keywords: Virtual Reality, Computer Vision, Object Detection, Public Street Environment, Simulation, Traffic, Pedestrian, Distribution, Labeling.

Abstract: Detecting moving objects in an urban environment is a challenging and widely explored problem in computer vision. This task requires huge amounts of data. Their obtaining and labeling is challenging. However the available datasets are not always fit the task. This work proposes a framework for synthesizing the train data based on 3D visualization of an urban environment using Unity 3D. Methods of mathematical statistics and distribution theory were used to build the background models of the framework. The framework, presented in the article, allows to simulate the real urban environment in an adjustable 3D virtual scene. It considers different environmental parameters amd makes possible to simulate the real behavior and physical characteristics of moving objects.

1 INTRODUCTION

There are an enormous amount of publically available datasets for the training of computer vision models. Nevertheless, there is a lack of data considering the specific parameters of scene or equipment, e.g. camera location, scene parameters, weather conditions, time of day, etc. Although, these parameters are critical for the dimensional based method developed in the previous works [1, 2]. The intrinsic and extrinsic parameters of the camera are significant for this algorithm.

Manual data collecting in real streets is time-consuming. There are streets with low traffic, hence the observation area is empty the majority of time. To provide data variety many different streets must be investigated. Moreover, of some rarely occuring events like car accidencs, appearance of jogging persons or disabled people must be gathered, it might take a while. Also, a real environment is pretty inconvenient for the development and debugging of object tracing algorithms [1]. In addition to the mentioned challenges, data labeling may take even longer time than data collection.

There are some approaches to synthesize the datasets by rendering the moving object on photo-background. However, these methods do not consider the effects which are important for background subtraction affecting the quality of the detection algorithms. Such possible effects are e.g. dynamic shadows, unexpected light reflections, or flares. Taking these effects into account makes the data more realistic.

A solution, proposed in this paper meets the above-mentioned requirements and allows the simulation of necessary conditions in 3D virtual reality. For this, Unity 3D engine provides a comprehensive toolset to recreate any kind of urban environment.

To make produced datasets representative in terms of a variety of moving objects the model creates pedestrians with different behavior and physiology based on data-driven statistical models. These models consider both, a variety in physiological features of real humans and their behavior changes depending on the environmental conditions and a time of day.

Statistical models are used for considering the changes in the behavior of vehicles during the day. It is also possible to vary the speed of the vehicle and produce ones with different colors.

The above considerations allow producing data that are very close to the real-world street conditions.

The remainder of this paper is structured as follows. Already existing methods are discussed in section 2. Section 3 describes the mathematical model

for generating traffic data. The proposed virtual environment simulatior is described in section 4. In section 5, results of this research along as future work suggestions are presented.

2 RELATED WORK

The idea of train data generation for object detection purposes may be addressed in different ways. In the work [3] a framework based on the Generative Adversarial Network (GAN) with multiple discriminators is proposed. It aims to synthesize realistic pedestrians on a given picture and learn the background context simultaneously. The framework includes the following components: generator G and two discriminators (D_b for background context learning and D_p for pedestrian classifying). The generator G replaces pedestrians on ground truth pictures with bound boxes, filled out with random noise and generates new pedestrians within that noise region. The discriminator D_b, learns to differentiate between real and synthesized pairs and forces the generator G to learn the background information like road, light condition in noise boxes. In the meanwhile, discriminator D_p learns to judge whether the pedestrian is a synthetic or real one. It leads to a smooth connection between the background and the synthetic pedestrian. After training, the generator can learn to generate photo-realistic pedestrians in the noise box regions and the locations of noise boxes are taken as the ground truth for detectors. Adding the Spatial Pyramid Pooling (SPP) layer in the discriminator D_p enables generation of pedestrians of different sizes.

The articles [4, 5] propose an efficient discriminative learning method that generates a spatially varying pedestrian appearance model that takes into account the perspective geometry of the scene. The method considers the surveillance setting where the following information is available: (1) intrinsic and extrinsic parameters of the static camera and (2) the geometrical layout of the scene, i.e., semantic labels for all the regions in the scene where a pedestrian could possibly appear (”pedestrian region”) and semantic labels for obstacles in the scene where a pedestrian could either be occluded or physically cannot be present. The area labeling performs manually. This obtained information is leveraged along with synthesized 3D pedestrian models to generate realistic simulations of the appearance of pedestrians for every location of the ”pedestrian region”. All artificial pedestrians are being rendered with respect to the camera parameters and the geometrical layout of the entire scene e.g., obstacles and occlusions. Consequently, these data are used to learn a smooth spatially-varying scene-specific discriminative appearance model for pedestrian detection.

The publication [6] proposes to replace or complement the real training data with augmented data, i.e. photo-realistic images comprised of virtual agents rendered onto a real image background. A sequence of real recorded images acquired from low-resolution vehicle-based cameras is used to reconstruct the surrounding 3D scene. Virtual pedestrians are being put in non-occluded positions and then animated in the reconstructed scene. Illumination is added to the scene to match the environment, and also simple geometry to cast and receive shadows. The bounding box for each virtual pedestrian is automatically generated using the alpha-mask obtained from rendering.

3 TRAFFIC GENERATION MODEL

The aim of a traffic generation model is to define a number of moving objects to be created by the simulator at each certain time period.

The straightforward approach is to generate data randomly [5]. However, this solution has one significant disadvantage: in this case, it is not possible to adjust the number of moving objects according to the real traffic situation in the particular street. In contrast, a data-driven approach, where the number of objects is generated by the predefined data-based function, enables simulating the traffic distribution for any particular street.

To obtain a data-based function one can either build a regression model, e.g. by using of neural networks [7] or use a probabilistic approach [8]. In the second case, the quantity of moving objects is to be considered as a random variable varying in the 24-hours time domain. The function itself is a probability density function (PDF) of a known distribution. This approach allows to directly find the expected number of objects from the histogram depending on the total number of objects while the regression model is a polynomial function of the n-th degree. The data preparation and modeling are described in the following subsections.

3.1 Data Preparation

The following moving objects are considered in the article: cars, motorcycles, trucks, buses, bicycles and pedestrians.

Three different data sources were considered to be the input in the modeling process.

- Traffic Counts - Hourly Classification Counts 2017 [9] - the dataset contains a set of records collected from different observation stations located in different roads of USA. Each record specifies vehicle count passed near a certain station on a certain date and within a certain hour. All vehicles are divided into several aggregated classes such as motorcycles, passenger cars, pickups and panel vans, buses, single-unit trucks, multi-unit tracks. From this dataset, only the data on car traffic was derived and used in further research.
- Bike Counts (Eco Counter) [10] - the dataset provides bicycle and pedestrian counts that were monitored at a number of locations in Edmonton, Canada. The data recording was carried out at 15-minute intervals. The part of the dataset related to bicycle traffic was derived to use in further research. Each record contains a vehicle count for two vehicle types (light or hard vehicles depending on their length). The part of the dataset related to bicycle traffic was derived to use in further research.
- Pedestrian Counting System – Monthly [11] - the dataset contains hourly pedestrian counts since 2009 from pedestrian sensor devices located across Melbourne, Australia. The data were aggregated from a variety of sources. The dataset was used in pedestrian traffic research.

Before modeling, the data were cleaned and averaged. During data cleaning, the duplicates and uncompleted rows were removed. Since the required model needs to describe the average behavior of the moving objects, the outliers in data were also removed. Although the second dataset includes both pedestrian and bicycle data, the last ones have many outliers and incomplete rows. For this reason, the third dataset was used for modeling pedestrian behavior.

After cleaning the data were averaged by the standard hour periods. After that, the whole observation history for each hour was also averaged. Finally, for each single observation point, the quantities of objects were obtained.

3.2 Data Modeling

Independently from the type of a moving object, the traffic is assumed to be normally distributed with the following PDF.

$$\phi_{\mu,\sigma^2}(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{1}$$

It is usually expected to see more moving objects in the streets in the morning and in the evening, than in the afternoon and at night, therefore one can assume the distribution to be bimodal [8].

To prove the above assumptions the traffic intensity histograms for averaged data were built. All the data were proven to be normally distributed. The assumption of bimodality was not confirmed because among the data there are cases where the distribution has a different number of peaks as shown in Figure 1.

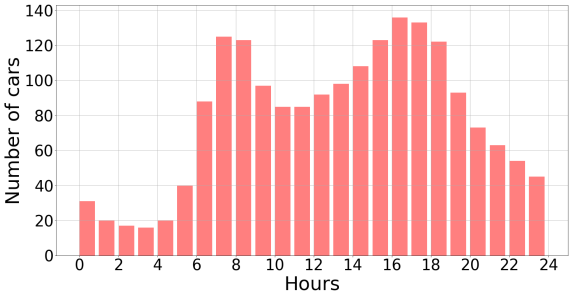


Figure 1: Averaged 24-hours traffic intensity histogram for individual cars.

The assumption of bimodality was not confirmed because among the data there are cases where the distribution have different number of modes. The above picture also illustrates, that the distribution is not purely bimodal, it has three local maxima. Therefore the general distribution is assumed to be multimodal. The equation 2 describes the distribution of traffic intensity I_t.

$$I_t = \sum_i A_i e^{-\frac{(t-t_i)^2}{s_i}}, \tag{2}$$

where i is a number of peaks on the histogram, t_i is a time where the i-th peak appears and A_i is a value of traffic intensity at the time t_i. This equation is an extension of that given in [8].

The above considerations allow using a Gaussian mixture model to estimate the parameters of PDF for the traffic intensity distribution. To do this the individual traffic intensities were recalculated into probabilities of observing the certain number of cars by the following formula:

$$P(I_t) = \frac{I_t}{\sum_i I_t} \tag{3}$$

The resulting PDF was obtained as a sum of PDFs for each $P(I_t)$ with an equal variance. The plot of individual PDFs for the cars is presented in Figure 2.

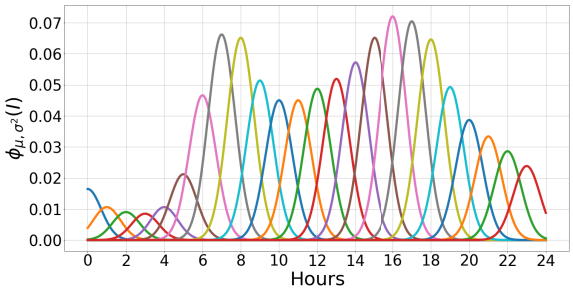


Figure 2: Plot of individual PDFs for the cars traffic intensities.

The resulting PDF for the car traffic intensity is presented in Figure 3. One can observe the first peak between 00:00 and 01:00 that will be lost when modeling with equation proposed in the paper [8].

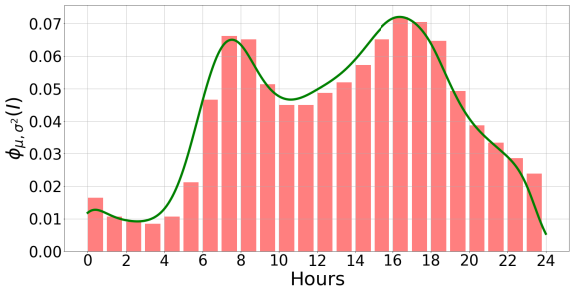


Figure 3: Histogram and resulting PDF curve for cars traffic intensities.

The resulting PDF for the traffic intensity of motorcycles is presented in Figure 4. The motorcycles are usually driving in the evening, that generally corresponds to real situation.

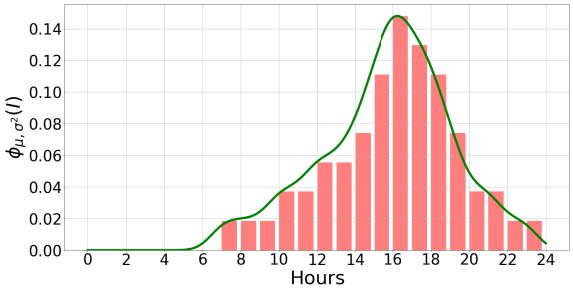


Figure 4: Histogram and resulting PDF curve for motorcycles traffic intensities.

The resulting PDF for the traffic intensity of trucks is presented in Figure 5. One can observe the highest peak at 08:00. After 11:00 the curve is closer to the uniform distribution than for cars.

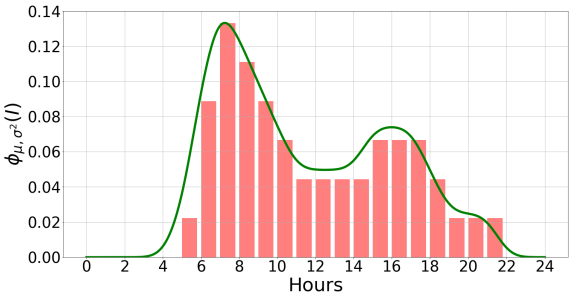


Figure 5: Histogram and resulting PDF curve for trucks traffic intensities.

The resulting PDF for the traffic intensity of buses is presented in Figure 6. One can observe only two segments on the histogram where the probability of observation is non-zero. The traffic of the buses is lower than that of other vehicles and that the buses use the schedule, while other objects move more randomly.

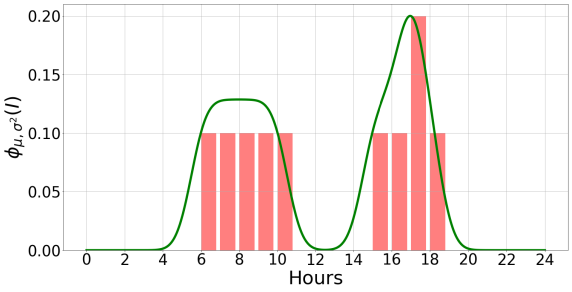


Figure 6: Histogram and resulting PDF curve for buses traffic intensities.

The resulting PDF for the traffic intensity of cars is presented in Figure 7. One can observe more extreme slopes in the PDF curve than in previous cases. It means, that cyclists are moving more randomly than other moving objects.

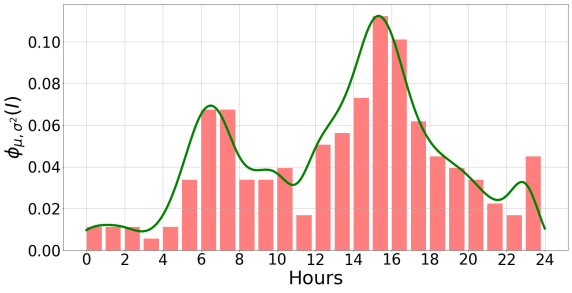


Figure 7: Histogram and resulting PDF curve for bicycles traffic intensities.

The resulting PDF for the traffic intensity of pedestrians is presented in Figure 8. One can observe multiple peaks on the histogram, therefore it might be

concluded, that the pedestrian traffic behavior is more complex than that for vehicles.

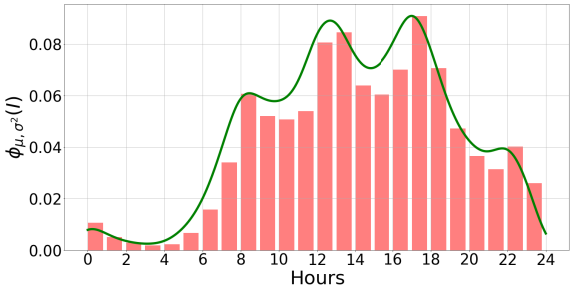


Figure 8: Histogram and resulting PDF curve for pedestrians traffic intensities.

To use the obtained model, one needs to define the total number of objects. Then the model will calculate the number of objects of a certain type at the given time point based on the PDF.

4 VIRTUAL ENVIRONMENT SIMULATOR

This section provides a detailed description of the testing infrastructure and software, which is used during the experiments.

The urban environment simulator's structure is shown in Figure 9, it consists of the following components:

1) Population Dataset - the dataset which contains the anthropometric parameters such as gender, height, width, weight, etc.
2) Generator of Anthropometric parameters - this component generates anthropometric parameters from the statistical model based on Population Dataset and passes the command to MakeHuman.
3) MakeHuman v1.2.0 (with Mass Produce plugin) - an open-source 3D computer graphics middleware designed for the prototyping of photorealistic humanoids. The example of generated models is shown in Figure 11.
4) Traffic History Dataset - a dataset which contains the data about the events that the object appears in the observation area [9, 10, 11].
5) Event Generator (EG) - a statistical model which generates the events that the object generated on step 3) will appear in the unity scene according to density data on step 4.
6) Unity 3D v2019.4.20f1 LTS - cross-platform 3d engine developed by Unity Technologies. It is used to generate the simulation scene of urban environments, animation of the moving object generated on step 3.The example of 3D scene is shown in Figure 12.
7) Labeled Dataset - the dataset obtained from 3D scene from step 6.

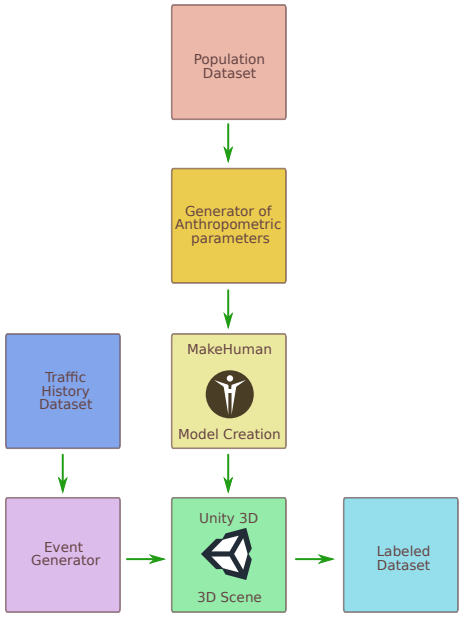


Figure 9: The components of the urban environment simulator.

4.1 Virtual Objects Generation

Pedestrian models are generated based on the anthropometrical studies of human diversity [12, 13]. The obtained statistical parameters which are shown in Table 1 are used to build a statistical model of the population which considers the height and weight of the individuals.

Table 1: Statistical parameters from population data.

	Male
μ_{height}	175.5 cm
σ_{height}	5.9 cm
μ_{weight}	74.49 kg
σ_{weight}	11.11 kg
	Female
μ_{height}	166 cm
σ_{height}	5.5 cm
μ_{weight}	60.3 kg
σ_{weight}	5.7 kg

From the statistical parameters, it is possible to make a generator of sets of anthropometrical parame-

ters for the models. The generator model can be described as bivariate normal distribution, which density function is shown in Figure 10.

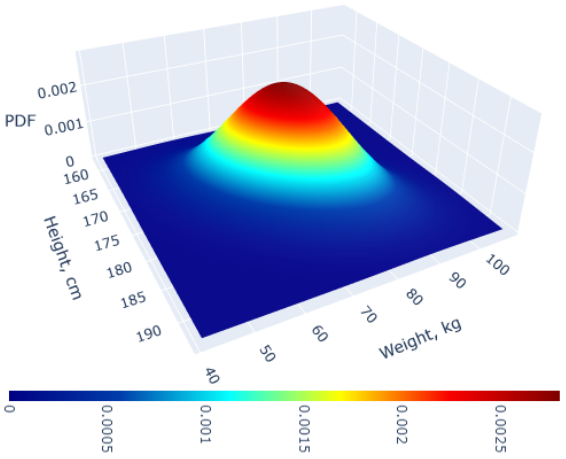


Figure 10: Bivariate normal distribution for height and weight for men.

The generated parameters are passed into MakeHuman software to produce 3D models. The models are shown in Figure 11.

Figure 11: Sample of pedestrians generated by MakeHuman.

The produced models will appear on the Unity 3D scene according to the scheduler from the event generator. The example of 3D scene is shown in Figure 12. The event generator is responsible for notification about the time and type of the object which will appear on the scene. It provides notification about several types of objects: pedestrians (male or female), cars, trucks, and cyclists. For vehicles, EG generates the parameters of color and type of vehicle (truck, bus, sedan, van, sports car, etc).

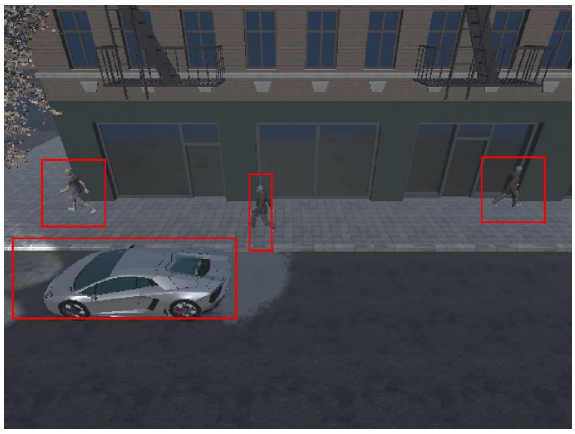


Figure 12: Daytime scene example.

Unity 3D application is responsible for urban environment visualization, illumination changes according to the time of day, animation of the objects, weather conditions on the scene, and generation of labeled screenshots with intrinsic camera parameters (FOV, height, and width of the matrix, IR filters, IR light 13, etc).

Figure 13: Nighttime scene with IR camera simulation.

5 CONCLUSIONS

The framework, presented in the article, allows to simulate the real urban environment in an adjustable 3D virtual scene. It makes it possible to simulate the real behavior and physical characteristics of both pedestrians and vehicles.

The models in the background of simulation are based on real data, which makes it possible to generate training data of high variety. To build them, the study of traffic behavior was carried out during the research. It was shown that traffic behavior generally has multimodal Gaussian distribution independently

on the type of moving object. Obtained PDFs allow using only the total number of objects per day to generate the objects at the scene. They also make the simulation close to the real situations and allow simulating custom events such as car accidents or public feasts based on user data.

Since the simulator is able to increase and decrease the speed of time, it is possible to simulate behavior during a large time period (e.g. month) faster than it is to be done in a real life. The simulator is also able to install different cameras at any instant point of the environment and adjust their parameters, which makes the data generation cheap and also allows to produce the data for the computer vision algorithms of many different types including dimensional-based ones. The parameters of illumination are also adjustable, which makes it possible to simulate night scenes or IR cameras.

The study of the model shows that virtual traffic simulation corresponds to the input data. This indicates that the simulation is consistent with the defined behavior of the objects.

The proposed framework could be instantly used for train data generation, however, there is room for improvement. In future the framework can be improved in the following ways:

Scene creation is a laborious and time-consuming task. Since the automatic scene generation should be considered. For example, a scene generation algorithm may take photos or videos as input data.

In addition to the above mentioned, it is possible to recreate the context of the scene. The scene context defines where the objects may be spawned and how they could behave.

It makes it possible to consider the traffic behavior as a periodic Gaussian process. The theoretical background of such a process is given in [14]. The statistical model is used for calculating the number of objects to be created on the scene at every instant time period. The statistics were obtained by averaging the real observations during the long time period. Finally, the average 24-hours histogram was built. However, under real-world the traffic behavior is cyclic, i.e. it is expected to find the same pattern at the same time each day. The traffic intensity variation limits can be investigated and then the model can be extended using the periodic Gaussian process.

REFERENCES

[1] I. Matveev, K. Karpov, A. Yurchenko, and E. Siemens, "The object tracking algorithm using dimensional based detection for public street environment," Eurasian Physical Technical Journal, vol. 17, pp. 123–127, Dec. 2020.

[2] I. Matveev, K. Karpov, I. Chmielewski, E. Siemens, and A. Yurchenko, "Fast object detection using dimensional based features for public street environments," Smart Cities, vol. 3, no. 1, pp. 93–111, 2020.

[3] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou, "Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond," arXiv preprint arXiv:1804.02047, 2018.

[4] W. Zhang, K. Wang, H. Qu, J. Zhao, and F.-Y. Wang, "Scene-specific pedestrian detection based on parallel vision," arXiv preprint arXiv:1712.08745, 2017.

[5] H. Hattori, V. N. Boddeti, K. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, Jun. 2015, pp. 3819–3827.

[6] J. Nilsson, P. Andersson, I. Y. Gu, and J. Fredriksson, "Pedestrian detection using augmented training data," in 2014 22nd International Conference on Pattern Recognition. IEEE, 2014, pp. 4548–4553.

[7] F. Moretti, S. Pizzuti, S. Panzieri, and M. Annunziato, "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling," Neurocomputing, vol. 167, pp. 3–7, 2015.

[8] L. Bartuška, V. Biba, and R. Kampf, "Modeling of daily traffic volumes on urban roads," 2016.

[9] Metropolitan Washington Council of Governments. Traffic Counts - Hourly Classification Counts 2017. Accessed Mar. 20, 2021. [Online]. Available: https://rtdc-mwcog.opendata.arcgis.com/datasets/fae4f4ebf99c45088adbfba504efd650

[10] Bike Counts (Eco Counter). City of Edmonton. Accessed Mar. 20, 2021. [Online]. Available: https://data.edmonton.ca/Monitoring-and-Data-Collection/Bike-Counts-Eco-Counter-/tq23-qn4m

[11] City of Melbourne Open Data Team. Pedestrian Counting System - Monthly (counts per hour). Accessed Mar. 20, 2021. [Online]. Available: https://data.melbourne.vic.gov.au/Transport/Pedestrian-Counting-System-Monthly-counts-per-hour/b2ak-trbp

[12] S. Buchmueller and U. Weidmann, "Parameters of pedestrians, pedestrian traffic and walking facilities," IVT Schriftenreihe, vol. 132, 2006.

[13] E. Brolin, "Anthropometric diversity and consideration of human capabilities," p. 101.

[14] N. HajiGhassemi and M. Deisenroth, "Analytic long-term forecasting with periodic gaussian processes," in Artificial Intelligence and Statistics. PMLR, 2014, pp. 303–311.

Models and Algorithms for Automatic Labelling of Unstructured Texts (Text Tagging)

Gyuzel Shakhmametova and Ilshat Ishmukhametov
Computer Science and Robotics Department, Ufa State Aviation Technical University, K. Marks Str. 12, 450008 Ufa, Russian Federation
shakhgouzel@mail.ru, mail@ishmukhamet.xyz

Keywords: Automatic Labelling of Texts, Unstructured Text, Text Tagging, Multilabel Classification, Keywords Extraction.

Abstract: The article discusses the task of automatic labelling of texts to improve the efficiency of processing unstructured text data. An overview of existing software products for solving the problem is given, showing the need to develop its own solution specialized in the processing of Russian-language texts. The problem of assigning labels is considered from a mathematical point of view as a problem of multilabel classification, with corresponding mathematical models analysed and described. Based on this, models, algorithms, and a software product for automatically assigning labels to texts have been developed. Numerical experiments were carried out that showed the universality of the method and the possibility of application both in non-specialized and specialized fields, in particular, for processing medical documents.

1 INTRODUCTION

Information systems are becoming more and more loaded and complex year after year, and the volume of information is growing at a tremendous rate. According to experts [1], by 2025 the volume of accumulated information will reach 175 zettabytes compared to 33 zettabytes in 2018 (1 zettabyte = 10^{12} gigabytes = 1 trillion gigabytes).

At the same time, most of the information is stored in unstructured form, mainly, as texts. In particular, among medical documents, structured data account for up to 20% of all available information [2].

In such circumstances, natural language processing tools are needed, since structuring the accumulated data significantly increases the efficiency of their use.

In working with text information, several large tasks may be distinguished, such as categorization task, identification of authorship, extraction of keywords and sentences, extraction of emotional context etc.

This article discusses the task of automatic assigning labels to texts, i.e., the so-called text tagging task, where each document from the corpus (set of texts) is mapped to tags (keywords, labels) from a certain set, helping to determine the content or purpose of the considered document.

Automatic labelling of texts is an urgent problem, since its implementation is necessary when solving problems in a variety of areas: in recommendation systems, electronic document management, in knowledge bases, etc. Text tagging tasks can vary significantly in specific cases and depend on the purpose, subject area, type, quantity and format of documents, language, etc. Accordingly, the methods of solving this problem can vary, making the choice of the appropriate method even more complicated.

The second part of the article presents related works in the field of text tagging; the third part is devoted to setting the task; the fourth part describes the proposed solution; results are discussed in part 5, and the sixth part contains the main conclusions.

2 RELATED WORKS

The relevance of the text tagging task contributes to its study by many researchers. As a result, several methods have been developed to solve it.

TextRank [3] is an adaptation of the PageRank [4] algorithm developed by Google to rank web pages. PageRank, in general, can be used to rank any group

of objects represented as a graph. TextRank converts the text into a graph and extracts keywords or sentences from it.

LDA (Dirichlet Latent Allocation) [5] uses the concept of hidden groups to determine text topics. It is assumed that each document may address several topics, and the appearance of a word in the text is related to one of these topics. In this way, labels can be assigned to the text corresponding to the themes of the package. A modification of LDA has also been developed to work with bigrams (two-word phrases) [6].

The RAKE (Rapid Automatic Keyword Extraction) [7] algorithm is based on the observation that keywords often consist of several words and, as a rule, do not include stop words (service parts of speech and the most used words). RAKE extracts phrases from the text using stop words as delimiters, and then counts estimates for them depending on how often words from these combinations are found in the document. Combinations with the highest scores are selected as keywords.

An algorithm has also been developed that simultaneously uses graph representation of text, LDA and RAKE [8].

In the application software market, there are many finished products for automatic tagging of texts.

- Dcipher Analytics [9] is a set of analytical tools for working with data, including text analysis. As the main areas of its work, Dcipher identifies Data Mining in the field of social media, automatic image processing, analysis of customer opinions, analysis of processes in the enterprise, etc. The platform provides the ability to build pipelines of data processing from a set of ready-made operations: importing data, collecting statistics, cleaning, and filtering data, training a model, etc. [9]
- MonkeyLearn [10] provides a service for automatic analysis of text data, such as automatic tagging, routing, and prioritization of requests from customers, analysis of user reviews, determination of customer mood, etc. The platform provides a comprehensive set of ready-made models from the built-in set [10].
- TwinWord [11] is a set of tools for developing texts based on the extraction and analysis of keywords, including analysis of the emotional connotation of the text, classification of texts, recommendation systems [11].

The main disadvantage of most such products is low flexibility in configuring the models to be used. Most often, only ready-made templates can be employed. Almost all existing solutions are close-ended, most of them have to be purchased for a fee, and, as a rule, they require an additional configuration. Additionally, there are practically no ready-made solutions focused on Russian-language texts.

All solutions available on the market are SaaS-based and provide APIs for integration into their own products.

A comparison of the software product characteristics described above is shown in Table 1.

Table 1: Product Comparison Results.

	Dcipher Analytics	Monkey Learn	TwinWord
Russian language	—	—	—
Documentation	—	Sufficient	Not sufficient
Price	From $3600 per year + trial	From $3600 per year + trial	Depends on the number of requests
Model Flexibility	High	Low	Low
Usability	Low	High	High

As a result of the analysis of the current state of research and software solutions in this field, it may be concluded that to fully support Russian-language texts and the ability to flexibly customize models, it is necessary to develop specific targeted solutions to the text tagging problem with the final implementation in the form of software.

3 PROBLEM DEFINITION

The product required for development should allow to create, edit, and delete text documents, label them, and merge them into collections. When new documents are added to existing collections, they must be automatically labelled (Figure 1).

Figure 1: Task Setting.

Assigning labels to texts fits the classification task definition: given a set of classified objects $X = x_1, \ldots, x_n$ (i.e. corpus of texts) and a set of classes $C = \{c_1, \ldots, c_m\}$ (i.e., label set). Objects and classes are related by a $\Phi: X \to C$ relationship.

Labelling in this formulation is suitable, for example, for categorizing documents. In this way, news may be classified ascribing each one of them into one of the categories such as «society», «politics», «sports», etc.

However, labelling in practice usually implies that documents are assigned more than one label. Returning to the above example, the same news item can simultaneously have labels «politics» and «economy».

Such a relationship between objects and classes may be described as follows (1):

$$\Phi: X \times C \to Y = \{0; 1\}^{|C|}, \quad (1)$$

where $y_{ij} = 1$ means that the object x_i belongs to the class c_j.

This describes the task of multilabel classification, i.e., such classification when one object can belong to several classes [12].

Thus, the task of text tagging is to construct a classifier (2):

$$\Phi': X \times C \to Y = \{0; 1\}^{|C|}, \quad (2)$$

where $X = \{x_1, \dots, x_n\}$ is a corpus of documents, $C = \{c_1, \dots, c_m\}$ is as set of labels, and $y_{ij} = 1$ means that the label c_j assigned to the document x_i, while Φ is the desired dependency between documents and labels.

4 SUGGESTED SOLUTION

There are several approaches to solving the problem of multilabel classification, which, in fact, are approaches with training potential.

- Reduction to binary classification [13]: its own binary classifier is built for each label separately, and the final set of labels for the document is created by determining which of these classifiers will give a positive result. It should be noted that such a solution loses some information, since correlations between labels are not taken into account.
- Reduction to multiclass classification [14]: in this case, label sets assigned to documents are perceived as separate classes. For example, for a set of two labels, $C = \{[0,0], [0,1], [1,0], [1,1]\}$ will be considered classes. A clear disadvantage of this approach is the large computational costs (exponential complexity) and the tendency to retrain, since not all possible sets of labels may occur in test data.
- Adaptation of multiclass classification methods is based on multilabel variations of the methods of kNN (ML-kNN [15]), decision trees (modification of the algorithm C4.5 [16]), and artificial neural networks (BP-MLL [17]).

As part of this work, adapted methods are considered because they take into account the correlation between assigned labels and are almost as computation-efficient as their multiclass counterparts.

In general, as with the other cases of using methods with training potential, the proposed solution can be divided into the following steps (Figure 2):

- preparation of data – the corpus of documents;
- feature extraction;
- model training;
- labelling new documents using a trained model.

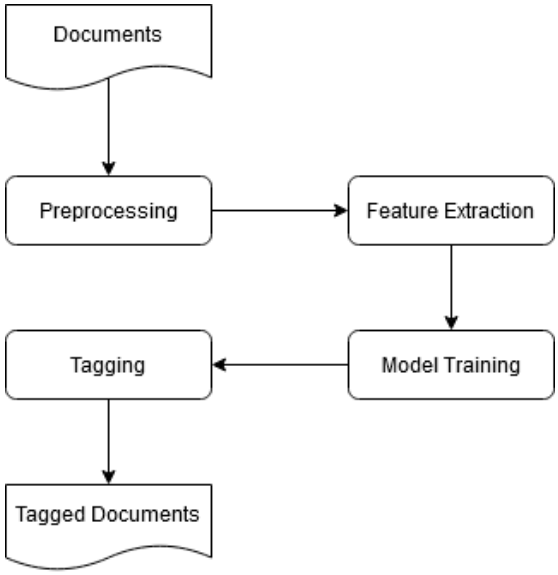


Figure 2: Algorithm of Text Tagging Task Solving as Multilabel Classification.

Let us take a closer look at each step.

4.1 Preprocessing

Among the obvious difficulties while solving this problem, one can distinguish the congestion of any text with service parts of speech: prepositions, conjunctions, particles, etc. They do not significantly affect the formation of the text topic but prevent the selection of keywords by frequency. Therefore, it is necessary to pre-process the text which is called normalization.

It is necessary to carry out lemmatization [18] – to bring all words to lemmas, their initial forms. For

example, for the word «stimulates», the lemma will be «stimulate», for «analysing» – «analyse».

An alternative to lemmatization is stemming – finding the basis (stem) of the word. For example, for the word «regulated», the stem will be «regul» which will allow to find such word forms as «regulate», «regulating», «regulation», etc.

The text normalization algorithm also includes the removal of punctuation signs and special characters, tokenization (division into word lists), the removal of stop words (using prepared in advance stop word lists) and the bringing of words to initial forms (using prepared in advance dictionaries).

Thus, the normalization of the text may happen according to the following algorithm (Figure 3):

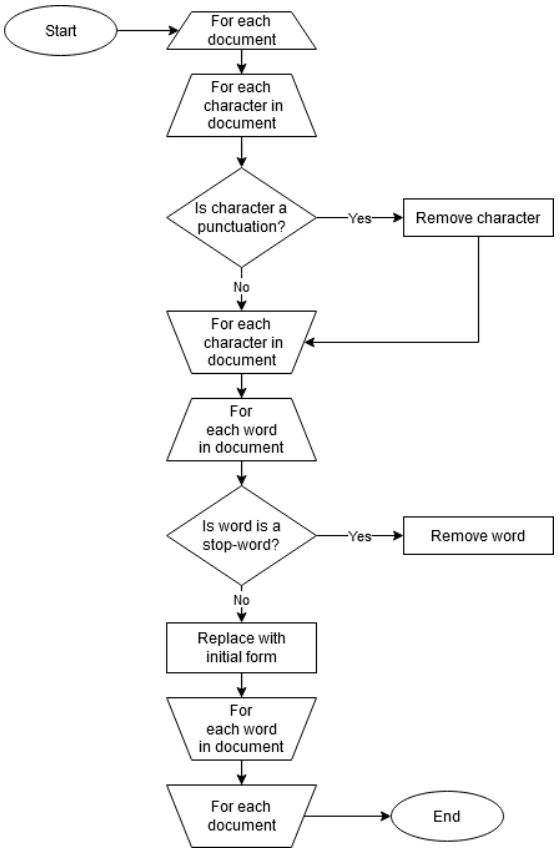


Figure 3: Text Corpus Preprocessing Algorithm.

4.2 Feature Extraction

TF-IDF model is used to extract features from pre-processed text corpus [19].

TF (Term Frequency) is the ratio for the number of occurrences of a given word to the total number of words in the text. The importance of the word is evaluated by the following (3):

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \quad (3)$$

where n_t is the number of occurrences of the word t in the document d, and the sum in the denominator is the total number of words in the document.

IDF (Inverse Document Frequency) is an inversion of the frequency with which a word occurs in body texts. Accounting for this indicator reduces the weight of words often used. Each word within a document collection has a value (4):

$$idf(t,D) = log\frac{|D|}{|d_i \in D : t \in d_i|} \quad (4)$$

where $|D|$ is the total number of documents in collection D, and the denominator is the number of documents in the collection in which the word t appears.

The TF-IDF measure is calculated as the product of the multipliers (5):

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (5)$$

4.3 Learning Algorithms

To train models in the work, the algorithms ML-kNN, Decision Tree and Random Forest were used.

4.3.1 ML-kNN

ML-kNN is a modification of the kNN method for multilabel classification. ML-kNN first determines the k nearest neighbours of the object. For those, it is already known which classes they belong to. Then, based on the maximum a posteriori estimation (MAP), it determines which labels to assign to the object in question.

The object x from the test sample with the label set Y_x is considered. Let $\overrightarrow{y_x}$ be the label vector for x, where the l-th component $\overrightarrow{y_x}(l)$ is 1 if the label l is assigned to the object x, that is, if $l \in Y_x$, and 0 otherwise. Let $N(x)$ be the set of indices k closest to x neighbors from the training sample.

Then, knowing a set of tags of these neighbours, we can define a membership counting vector (6):

$$\overrightarrow{C_x}(l) = \sum_{a \in N(x)} \overrightarrow{y_a}(l), l \in Y \quad (6)$$

That is, this vector counts the number of neighbours labelled l.

Let the event H_1^l indicate that the object in question has the label l, and H_0^l indicates that it does not. Let the event $E_j^l, j \in \{0, \dots, k\}$ show that among the k closest neighbours of object x there are exactly j objects with the label l.

Then, based on the membership calculation vector and using the maximum a posteriori estimation, it is possible to determine the vector of labels $\overrightarrow{y_x}$ for $l \in Y$ in this way:

$$\overrightarrow{y_x}(l) = argmax_{b \in \{0,1\}} P\left(H_b^l \left| E^l_{\overrightarrow{C_x}(l)} \right.\right) \quad (7)$$

Using Bayes' theorem, this expression can be brought to the form:

$$\overrightarrow{y_x}(l) = argmax_{b\in\{0,1\}} \frac{P\left(H_b^l\right)\cdot P\left(E^l_{\overrightarrow{C_x}(l)}\middle|H_b^l\right)}{P\left(E^l_{\overrightarrow{C_x}(l)}\right)} \quad (8)$$

Probability $P\left(E^l_{\overrightarrow{C_x}(l)}\right) = 1$, therefore:

$$\overrightarrow{y_x}(l) = argmax_{b\in\{0,1\}} P\left(H_b^l\right)\cdot P\left(E^l_{\overrightarrow{C_x}(l)}\middle|H_b^l\right) \quad (9)$$

Probabilities $P\left(H_b^l\right)$ and $P\left(E^l_{\overrightarrow{C_x}(l)}\middle|H_b^l\right)$ can be calculated on a training sample.

4.3.2 Decision Tree

The C4.5 algorithm uses the concepts of entropy and information gain criteria to determine an attribute for better splitting a training set into a tree.

Let be given a training set S containing m attributes and n objects belonging to k classes. The tree is built from the root node to the leaves, that is, from top to bottom.

In the first step, an empty tree is built, consisting only of a root that includes the entire set S.

Next, the root is split into subsets and child nodes are defined. To do this, one of the attributes is selected and a rule is formed that breaks the set of objects into p subsets, where p is the number of unique values of the selected attribute. The procedure is then repeated for each of the received subsets and the child nodes. The procedure continues until the stop condition is reached.

Let $N\left(c_j, S\right)$ denote the number of objects of class c_j in the set S, and $N(S)$ denote the total number of examples in the set S. Then, the relative frequency of class c_j in the set S can be determined:

$$p\left(c_j\right) = \frac{N(c_j,S)}{N(S)} \quad (10)$$

Variable

$$H(S) = -\sum_{i=1}^{k}[p(c_i)\cdot log(p(c_i))] \quad (11)$$

is an entropy of a set S and shows the average amount of information needed to determine an object class from that set.

After dividing the set by attribute A, this estimate can be written as:

$$H_A(S) = \sum_{i=1}^{k}[p(c_i)\cdot H(S_i)], \quad (12)$$

where S_i is the i-th node that was obtained during the partition. Then the best split attribute can be selected using the information gain criterion:

$$IG(A) = H(S) - H_A(S) \quad (13)$$

For partitioning, an attribute is selected, for which the gain in information is the greatest.

If an empty node is formed during the split process, it becomes a sheet, and a class which more often was met among objects of the parent node G is associated with it.

The above formulas apply to discrete attributes. In the case of a continuous attribute having n different values, the set of its values is divided into n subsets using $(n-1)$ threshold values. Using the information gain criterion, the threshold value that gives the largest information gain is selected.

To use the C4.5 algorithm for multilabel classification, the entropy count is changed as follows:

$$H'(S) = -\sum_{i=1}^{k}(p_i + q_i), \quad (14)$$

where $p_i = p(c_i)\cdot \log p(c_i)$, $q_i = q(c_i)\cdot \log q(c_i)$, $q(c_i) = 1 - p(c_i)$.

4.3.3 Random Forest

The standard implementation of the random forest method with trees described in paragraph 4.3.2 was used.

For a training sample S of size N with M attributes, a random forest is described as:

$$\{h(x,\Theta_k), k = 1, \dots\}, \quad (15)$$

where $h(x,\Theta_k)$ is a separate tree built on a subset Θ_k of the training set.

The forest building algorithm includes the following steps:

1) From the training sample S, a subset Θ_k of size N is randomly generated with repetitions: some objects will be included more than once, some will not be included at all.
2) On the obtained sub-sample, a tree $h(x,\Theta_k)$ is built using not the entire set of features, but only m randomly selected.

Thus, several trees are built. Their optimal number is selected to minimize classification errors on the test sample.

5 RESULTS

The above algorithms were used to implement the software TextTagger [20], designed to automatically assign labels to texts.

The machine learning module is implemented using Python and the main libraries for machine learning and NLP (nltk, sclearn, etc). The business logic module is implemented on the .NetCore platform, while the web client is developed using the Vue framework.

During operation, the user generates collections consisting of text documents. The first added documents are labelled manually, and later this initial classification will be used to train a model.

After adding several documents to the collection, one of the above models (ML-kNN, Decision Tree or Random Forest) can be trained. The user selects the model manually.

When a new document is added to a collection with a trained model, the system automatically prompts to assign the most appropriate labels to it.

Figure 4 and Figure 5 show the main steps of working with the system using the example of accumulating a collection of medical documents.

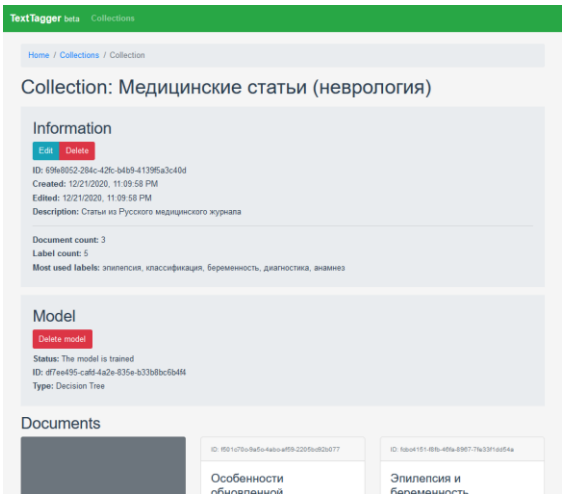


Figure 4: Collection Overview Page.

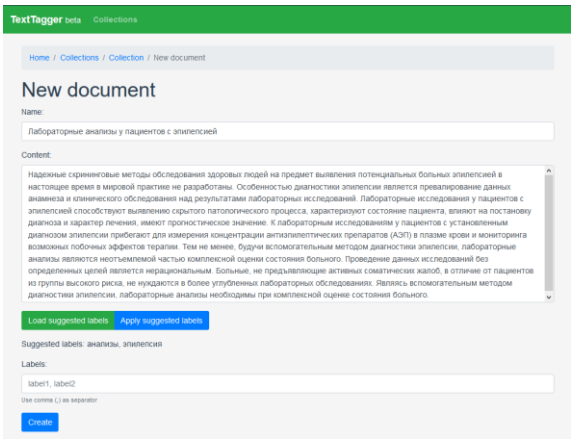


Figure 5: Adding a New Document with the System Proposed Labels.

Based on experimental data, the quality of the algorithms was analysed. Accuracy (A), precision (P), recall (R) and F1-score ($F1$) are used as quality metrics [21].

$$A = \frac{1}{n_i}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}, \quad (16)$$

$$P = \frac{1}{n_i}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|}, \quad (17)$$

$$R = \frac{1}{n_i}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|}, \quad (18)$$

$$F = \frac{1}{n_i}\sum_{i=1}^{n}\frac{2\cdot|Y_i \cap Z_i|}{|Y_i|+|Z_i|}, \quad (19)$$

where $Y_i = \{0;1\}^k$ are the labels assigned to the document x_i (actual) and $Z_i = \{0;1\}^k$ are the labels predicted by the model.

The closer the scores are to 1, the higher the quality of the model.

Several prepared in advance and labelled datasets from different subject areas were used to conduct quality analysis of algorithms:

Set A: abstracts from literature texts; labels are age groups to which texts are oriented (for children, for adults, for children and adults at the same time). This dataset consists of 75 documents, the labels are evenly distributed.

Set B: synopsis of films; tags are their genres (drama, comedy, etc). This dataset consists of 50 documents, the labels are unevenly distributed.

Set C: annotations to articles from the Russian Medical Journal; labels are medical concepts referred to in the articles (cognitive impairment, therapy, etc. This dataset consists of 56 documents, the labels are unevenly distributed.

The quality metrics for the Decision Tree, Random Forest and ML-kNN algorithms are shown in Table 2.

Table 2: Quality Metrics.

Method	Data set	A	P	R	F1
DT	A	0.94	0.97	0.97	0.96
	B	0.27	0.50	0.41	0.38
	C	0.36	0.46	0.42	0.41
RF	A	0.78	0.86	0.80	0.81
	B	0.25	0.65	0.27	0.36
	C	0.23	0.27	0.26	0.26
ML-kNN	A	0.94	0.94	1.0	0.96
	B	0.32	0.56	0.38	0.42
	C	0.60	0.69	0.64	0.64

The quality of the models used depends significantly on the specific training data, especially on the uniformity of the distribution by labels. However, overall quality allows models to be used to solve some of the practical problems or as a basis for further development and improvement.

Random Forest (metric F1 on average in three datasets 0.47) generally showed itself worse than Decision Tree (0.58), and Multilabel KNN (0.67) showed the best result.

In the future, the quality can be increased, for example, due to heuristics. On different types of data, different models work with different efficiencies. It makes sense to identify the appropriate patterns and select the most suitable models for specific data sets, based on the uniformity of the data, the volume of texts and other characteristics of the cases.

Even though the solution is initially aimed at working with Russian-speaking corpuses, it can be applied to other languages as well, in particular, to English.

Considering further improvement of quality, the product can be used in decision support systems or be integrated into knowledge bases. It seems promising to use it to mark-up texts of medical topics for highlighting the main concepts in them, which can be used to structure texts and further process them. One of the methods of application planned by the authors is to highlight keywords in the texts of clinical recommendations to further structure them for use in clinical decision support systems.

6 CONCLUSIONS

The problem of automatically assigning labels to texts is current, and its solution is in demand in many tasks of processing unstructured texts. The analysis of methods and existing software products for the text tagging task showed the lack of ready-made tools for processing Russian-language texts and the need to solve the text tagging problem with the final implementation in the form of software.

As a solution, it is proposed to consider text tagging as a problem of multilabel classification. A comparison of the known methods of solving the problem of multilabel classification was made, and, subsequently, the following methods with training were selected: ML-kNN, Decision Tree (ML-C4.5), and Random Forest. Additionally, TF-IDF method was included to extract features.

Based on the selected models, algorithms have been developed for automatically assigning labels to texts, and then implemented as the TextTagger software product.

The computational experiment showed a fairly high efficiency of the developed models for ML-kNN (metric F1-Measure 0.67) and the average for Decision Tree (0.58) and Random Forest (0.47), which indicates the possibility of their use in practice. Further quality improvement is possible by refining the process of normalizing texts and introducing heuristics to select the best possible model for a specific data set.

The developed software product is universal, applicable in various subject areas for processing texts in Russian, and applicable to other languages.

ACKNOWLEDGMENTS

The reported study was funded by RFBR according to the research projects No 19-07-00780, 19-07-00709.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, “The Digitization of the World – From Edge to Core,” IDC white paper, November 2018, Doc# US44413318.

[2] A. M. Nancy and R. Maheswari, "Review on unstructured data in medical data," Journal of Critical Reviews, 2020, pp. 2202-2208, doi: 10.31838/jcr.07.13.342.

[3] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” EMNLP, 2004.

[4] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Comput. Networks, vol. 30, 1998, pp. 107-117.

[5] S. Tasci and T. Güngör, “LDA-based keyword selection in text categorization,” 24th International Symposium on Computer and Information Sciences, 2009, pp. 230-235.

[6] A. Sedova and O. Mitrofanova, “Topic Modelling of Russian Texts based on Lemmata and Lexical Constructions,” Saint-Petersburg State University, 2017.

[7] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic Keyword Extraction from Individual Documents,” 2010.

[8] M. Thushara, M. Krishnapriya, and S. N. Sangeetha, “A model for auto-tagging of research papers based on keyphrase extraction methods,” International Conference on Advances in Computing, Communications and Informatics, 2017, pp. 1695-1700.

[9] Dcipher Analytics official web-site [Online]. Available: http://www.dcipheranalytics.com.

[10] MonkeyLearn official web-site [Online]. Available: https://monkeylearn.com/.

[11] TwinWord official web-site [Online]. Available: https://www.twinword.com/.

[12] J. Read , B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” Machine Learning, vol. 85, 2011, pp. 333-359.

[13] S. Godbole and S. Sarawagi, “Discriminative Methods for Multi-labeled Classification,” PAKDD, 2004.

[14] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," Int. J. Data Warehous, Min. 3, 2007, pp. 1-13.

[15] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit, vol. 40, 2007, pp. 2038-2048.

[16] A. Clare and R. King, "Knowledge Discovery in Multi-label Phenotype Data," PKDD, 2001.

[17] M. Zhang and Z. Zhou, "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization," IEEE Transactions on Knowledge and Data Engineering, vol. 18, 2006, pp. 1338-1351.

[18] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," 2014.

[19] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based Framework for Text Categorization," Procedia Engineering, vol. 69, 2014, pp. 1356-1364.

[20] TextTagger online demo [Online]. Available: https://texttagger.ishmukhamet.xyz.

[21] A. Luque, A. Carrasco, A. Martín, and A. D. L. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., vol 91, 2019, pp. 216-231.

Technology of Computer Monitoring of the Quality of Educational Process

Tatiana Monastyrskaya[1], Alexey Poletaikin[1, 2], Julia Shevtsova[1] and Elena Melekhina[3]
[1] *Siberian State University of Telecommunications and Information Science, Kirova Str. 86, 630102 Novosibirsk, Russian Federation*
[2] *Kuban State University, Stavropol'skaya Str. 149, 350040 Krasnodar, Russian Federation*
[3]*Novosibirsk State Technical University, K. Marksa avenue 20, 630073 Novosibirsk, Russian Federation*
t.monastyrskaya@mail.ru, alex.poletaykin@gmail.com, shevcova_yuliya@mail.ru, melexina@corp.nstu.ru

Keywords: Quality of Education, Sociological Monitoring, Computer Monitoring, Risk Thermometer, Key Risk Indicators, Fuzzy Composition, Management Decision.

Abstract: This research aimed at developing operational assessment tool to minimize the university risk background with the purpose to raise the quality of the educational process. The original mathematical approach is proposed as a means to solve the problem of assuring the quality of education. The method of modified risk thermometer and binary fuzzy relations composition were used as the basic methods of sociological monitoring data analysis to measure the satisfaction of students with educational process. The method of modified risk thermometer identifies the risk background of the educational process, defined by the Key Risk Indicators. The method of fuzzy analysis allows to consider and minimize the existing uncertainty of the educational process and risk background. It is shown on the example that if the university risk background is of high degree, it necessitates taking the complex of management decisions to improve the situation with the risk background. The theoretical significance of the research is in development of the methodology of educational computer monitoring. The application of this methodology raises satisfaction of students and teachers with educational process, objectivity of management decisions and their implementation into educational process in order to normalize the risk temperature, which is the practical significance of the research. The degree of this condition corresponding to the normal one is defined at the next stage and needs taking further management decisions. The described methodology is a universal and efficient tool to revaluate the activity of not only universities but also of any company at risk as well as to organize the process of risk management in social and economic systems.

1 INTRODUCTION

The sociological research into universities' competition for top positioning in global and national rankings demonstrates the growing demand for the ways to monitor the university performance [1, 2, 3, 4]. According to B. Williamson, the findings of the recent sociological research conducted in the United Kingdom raise 'two critical points': the traditional judgement made by experts and professionals is substituted with numerical data, and the people's understanding of the notions ,good university' and ,good course' is changing due to the rankings' results [5]. So, literature suggests that most attention has focused on monitoring as an instrument to improve the performance of an organisation, that Lucas H., Greely M. and Roelen K. define as 'higher frequency data collection or reporting, often using information and communication technologies, to strengthen current programme performance or to inform policy and the practice (design, scale and scope) of future service delivery' [6]. Any monitoring system aims to deal with stability and availability [7], that is why, it needs to be reliable and efficient.

One of the monitored parameters to measure the university's performance is the quality of the educational process. It needs monitoring not only for measuring progress and growth but also for negative trends and risks, which is understood as ,the effect of uncertainty on objectives resulted in a deviation from the expected — positive and/or negative and is often expressed in terms of a combination of the consequences of an event (including changes in

circumstances or knowledge) and the associated likelihood of occurrence' [8]. Birkinshaw J., and Jenkins H. define risk as 'the potentially negative impact arising from a future event <…> that can be calculated as a product of the probability of the future event happening and the scale of loss associated with that event'[9].

In order to measure and manage the risks of the educational process the qualitative and quantitative criteria need to be identified [10]. These criteria, which could serve as an operational risk management tool, are called key risk indicators. Young J. underlines that a risk indicator 'becomes key when it tracks a risk exposure, which could have a major influence on the organisation'[11].

Multilevel monitoring is an efficient social technology for managing the quality of professional education [12] in conditions of dynamically changing educational standards, developing new technologies, resources and forms of teaching resulted from globalization and internationalization. The university authorities also change forms and methods of management to make the university competitive. They need new ways of information collection and processing. The ongoing innovations in dynamic assessment of changes make social monitoring the efficient tool for universities' top and middle management. Monitoring helps not only detect changes, but also evaluate the results of managerial decisions [13].

The latest Federal state educational standards in Russia call for internal and external assessment of the quality of education [4]. Monitoring and operational one-time research allow to work out the methodology of development and objective assessment of the quality of teaching and learning within any university professional program on the principles of competence-based approach [15, 16]. It necessitates the research and development of mathematical model of computer monitoring of education with the purpose to minimize educational risks.

The sociological research conducted at Siberian State University of Telecommunications and Information Science defined the quality of education as a social category which has the following characteristics [17]:

- it defines conditions and efficiency of educational process in society, its meeting the needs and expectations of the society for learners' social, personal and professional competences;

it is measured by the complex of indicators characterizing various parameters of university performance, which provide the development of learners' competences: curriculum, forms and methods of teaching, facilities, staff. The data were collected by the risk thermometer method, which is considered to be one of the effective tools to measure risk background of social and economic system and make the first approximation to identify organizational risk background.

The monitoring literature analysis demonstrates that various methodological approaches are used for monitoring consumers' satisfaction with goods and services. Such approaches, methods and techniques include «SERVQUAL», «SERVPERF», «INDSERV», CSM, weighted estimate method, discrepancies analysis method [18]. However, with the aim to analyze not only students' and teachers' satisfaction with educational process, but also its riskiness, we have chosen the risk thermometer method as a model-measuring approach of risk management. Without well-developed corporate culture of risk management any organization uses 'primitive' methods of risk management like risk mapping, risk calculator, risk radar, and risk thermometer. The risk thermometer method allows defining the risk background of an organization at a first approximation. Moreover, its application does not require from the user any special risk management knowledge and skills, which might be considered the advantage of the method. This is due to the fact that the risk thermometer method is based on carrying survey data to the integral indicator, which can be interpreted as the risk temperature of a company, with survey questionnaire items implicitly indicating company's risks and being not specific for risk management.

2 FORMAL DESCRIPTION OF THE RISK THERMOMETER

Being the formalization of a survey procedure [19], the risk thermometer method leads the statistically processed survey results from a questionnaire to the integral indicator:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{m} k_{ij}^{l} x_{ij}, \; l = \overline{1, p},$$

where x_{ij} – variable of j-th respondent's answer to i-th item of the questionnaire: $x_{ij} = 1$, if respondent gave i-th answer to j-th question, $x_{ij} = 0$ – vice versa; k_{ij}^{l} – risk value of j-th answer to i-th question; l – index of the object risk condition, p – number of such conditions.

Risk coefficients k^l are expertly set and serve as a norm coefficient leading integral results to temperature indicators: $k^l = \frac{T_l}{n},\ l = \overline{1, p}.$ For example, when analyzing if some system meets the requirements, it is possible to see the following temperature conditions T_l:

- normal $T_1 = 36{,}6\,°C$ – full compliance;
- fever $T_2 = 38\,°C$ – partial compliance;
- hazardous $T_3 = 42\,°C$ – full inadequacy.

As far as the research of a complex system requires a complicated questionnaire structure its items are grouped according to some value-based criteria (block 1). The received groups are considered to be the target factors, rational and purposeful actions aimed at adjusting the researched system to normal conditions meeting the standard requirements (see Table 1). The baseline study [17] worked out the students' questionnaire including 79 items addressed to the learners and aimed to receive qualitative assessment of satisfaction according to five-grade scale shown in Table 1.

Table1: Risk thermometer to monitor the quality of educational process.

Satisfaction assessment		Risk temperature	
Score	Verbal	°C	Verbal
5	To a full degree	36,6	Normal
4	To a degree	37,2	Subfebrile
3	To a moderate degree	38,0	Feverish
2	To some degree	39,5	Critical
1	To no degree	42,0	Hazardous

The questionnaire items were organized into eight groups (see Table 2), which can be considered as satisfaction indicators characterizing its specific aspects and correlating with the goals of educational program risk temperature measurement. In fact, the questionnaire reflects the organizational values being at risk and allows to analyze if they meet the requirements of the Federal State Educational Standard.

The questionnaire serves as baseline data for the risk-management model. The group of questions to the faculty is interpreted as value-based criteria, rational and purposeful actions aimed at adjusting the researched system to normal conditions meeting the standard requirements.

Table 2: Groups of program satisfaction indicators.

Groups of questions in students' questionnaire	Objects of risk	Complex satisfaction factor
1 Degree of your satisfaction with learning	Processes	Satisfaction with different learning activities
2 Degree of your satisfaction with teaching	Personnel	Satisfaction with teachers' work
3 Degree of your satisfaction with organization of the learning process	Processes	Satisfaction with organization of the learning process
4 Degree of your satisfaction with university facilities	Systems	Satisfaction with university facilities
5 Degree of your satisfaction with the quality of university services	Services	Satisfaction with the quality of university services
6 Degree of your satisfaction with extracurricular activity	Reputation	Satisfaction with extracurricular activity
7 Degree of your satisfaction with information support of curricular and extracurricular processes	Processes	Satisfaction with information support of curricular and extracurricular processes
8 Degree of your satisfaction with studying at university in general	Reputation	Satisfaction with studying at university in general

The differential (partial) risk temperature (T^s) with reference to aggregate satisfactory data group (its efficiency q=8) is defined in the following way:

$$T^s = \frac{n}{n^s} \sum_{i=1}^{n^s} \sum_{j=1}^{m} k_{ij}^l x_{ij},\ \ l = \overline{1, p},\ \ s = \overline{1, q}. \qquad (1)$$

The calculated temperature data are to some degree fuzzily uncertain due to the specificity of the risk thermometer method. Uncertain temperature parameters of the risk thermometer cannot ensure high validity of assessment. Even border risk thermometer values in Table 1 cannot be considered fairly reliable because they represent experts' assessment influenced by subjectivity. In terms of risk thermometer method the researched object behavior within the interval between the borders remains uncertain to some degree. It is advisable to formalize such regions of uncertainty by fuzzy numbers, and conduct temperature data processing in a fuzzy form.

Multiple studies, including research into education [20-24], demonstrate the efficiency of fuzzy logics for solving the problems of educational management. The most complete set of the fuzzy data analysis methods and their application to management is represented in [20]. Laal M. proves the appropriateness of the fuzzy educational data analysis and application of the fuzzy propositions for the formalized assessment of results [22].

That is why, their further processing is necessary to fulfill by means of fuzzification on the basis of membership function. In this research the membership function is defined by the experts' opinion (authors of the article) with reference to linguistic scale (Figure1). For the usability of the model and with consideration of fuzzy mode of educational control [22-24], in particular, the research of the similar fuzzy model [24], z-shaped and s-shaped membership functions were used for border linguistic terms and trapezium-shaped membership functions within the interval of uncertainty. Taking into consideration the inviability of any system (both real and artificial) at hazardous temperatures it was decided to confine the MF to four levels, understanding them as linguistic terms of fuzzification procedure: N – Normal, AN – Above the Normal, H – High, Cr – Critical.

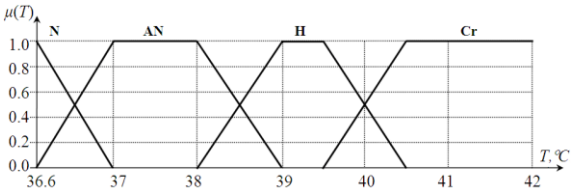


Figure 1: Membership functions of linguistic variable "Fuzzy risk temperature".

3 THE FUZZY TECHNOLOGY OF MONITORING THE RISK BACKGROUND OF A UNIVERSITY EDUCATIONAL PROCESS

The risk background monitoring was carried out with the system of aggregate data called key risk indicators [25, 26]. In this research, they are target factors $c_j \in C$, $j = \overline{1,q}$, which are indicators to identify characteristic risk events $r_i \in R$, $i = \overline{1,n_r}$. The matrix of $n_r \times q$ size defines the correlation of the target factors and risks, and can be considered the binary fuzzy ratio $RC = \left\{ \left\langle r_i, c_j \right\rangle, \mu_{RC} \left\langle r_i, c_j \right\rangle \right\}$. The next stage of the risk management process is the execution of the composition binary fuzzy ratio:

$$R_T = RC \otimes C_T , \quad (2)$$

where RC-binary fuzzy ratio, containing reference fuzzy data about risk matching key risk indicators (see Table 3), C_T - binary fuzzy ratio, contained fuzzified assessment of the differential risk temperature (1) for key risk indicator $c_j \in C$ by assessment of the level $l_k \in TL$ ($k = \overline{1,4}$) in the form of linguistic terms in accordance with membership function 'Fuzzy risk temperature' graphically represented in Figure 1.

Table 3: Frequency characteristics of risk identification.

Identified risks	Key risk indicator category (see Table 2)							
	1	2	3	4	5	6	7	8
1 Risk of knowledge obsolence	0,70	0,50	0,00	0,14	0,00	0,00	0,83	0,45
2 Mismatching of the stakeholders interests	0,53	0,50	0,00	0,14	0,00	0,00	0,67	0,36
3 Technical system malfunctioning	0,23	0,38	0,75	1,00	0,00	0,20	0,83	0,64
4 Risk of the key personnel dependence	0,27	0,25	0,75	1,00	1,00	1,00	1,00	0,45
5 Personnel depletion	0,17	0,25	0,00	0,14	0,00	0,40	0,17	0,64
6 Stagnation of research	0,60	0,50	0,00	0,00	0,00	0,00	0,33	0,73
7 Devaluation of personnel creativity	0,30	0,63	0,00	0,57	0,11	0,00	0,33	0,45
8 Lack of identity and uniqueness	0,37	0,88	0,75	1,00	1,00	1,00	1,00	1,00

The explicit quantitative assessment of *i*-th risk r_i can be received by defuzzification of the method of center of gravity for the one-point set:

$$P_{r_i} = \frac{1}{\sum_{k=1}^{4} r_{ik}} \sum_{k=1}^{4} (r_{ik} \cdot b_{ik}),$$

where r_{ik} is the element of binary fuzzy ratio R_T corresponding *k*-th term $k = \overline{1,4}$; b_{ik} - the explicit value of the corresponding element of baseline set of factors *TF*, defined on the basis of membership function 'Fuzzy risk assessment' graphically represented in Figure 2 according to the scale in Table 4.

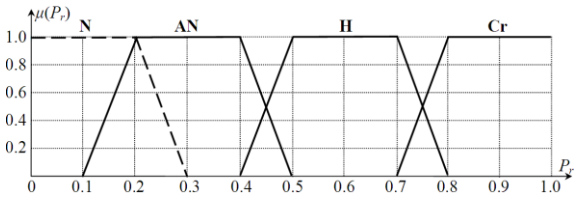


Figure 2: Membership functions of linguistic variable "Fuzzy risk assessment".

Table 4: The scale for fuzzy assessment risk probability [15].

Notion	Rank	Interpretation of risk probability	
N	1	Remote probability	$P_r < 0.20$
AN	2	Mean probability	$P_r \in [0.20, 0.50)$
H	3	High probability	$P_r \in [0.50, 0.75)$
Cr	4	Very high probability	$P_r \geq 0.75$

4 THE RISK ASSESSMENT ON THE BASIS OF THE QUESTIONNAIRE DATA

There were 392 respondents in the survey conducted at Siberian State University of Telecommunications and Information Science by means of questionnaire mentioned in part 2. The total number of data received in the form of qualitative assessments given in answers to 79 questions amounted to 30 thousand elements of data. Moreover, the questionnaire traditionally used as a part of university accreditation in the Russian federation [7] was given to teachers, because teachers as well as students are the key players in educational systems and their satisfaction characterizes the quality of the educational process of the university. The distribution of the received assessment according to the scale is represented in Figure 3.

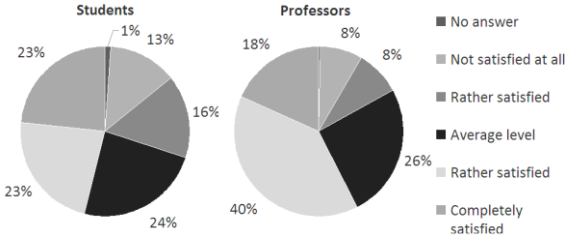


Figure 3: The structure of satisfaction assessment by the university students and professors.

According to the method of transformation assessment grades into temperature data shown in Table 1 and (1), the following risk temperature assessments of the university educational process given by students and teachers were received (Figure 4). It is worth noting that the received data demonstrate high consistency, for example, students' variation coefficient is 1,78%, and teachers' – 2,13%. In students' graph: 1 is studying; 2 – teaching; 3 – organization of educational process; 4 – facilities; 5 – quality of services; 6 – extracurricular activities; 7 – information support; 8 – integral risk temperature. In teachers' graph 1 is facilities and resources; 2 – educational process organization; 3 – working conditions; 4 – integral risk temperature.

The risk temperature, calculated by the (1) for every key risk indicator is shown in Figure 4 where its fuzzified values correlated with the MF defined by the experts (Figure 1).

The given example demonstrates that the identified risk background requires taking management decisions to reduce the risk background up to the optimal level (N - according to the scale in Table 5). As a result of such decisions the key risk indicator will become normal.

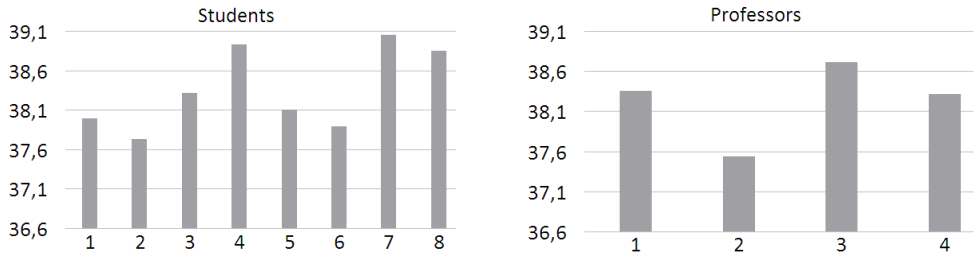


Figure 4: Risk temperature of the university students' and teachers' satisfaction.

Table 5: Temperature characteristics of satisfaction indicators.

Indicators group (key risk indicator category)	Indicators number	Risk temperature				
		T, °C	N	AN	H	Cr
Satisfaction with different learning activities	30	37,99	0,00	1,00	0,00	0,00
Satisfaction with teachers' work	7	37,73	0,00	1,00	0,00	0,00
Satisfaction with organization of the learning process	4	38,31	0,00	0,69	0,62	0,00
Satisfaction with university facilities	7	38,93	0,00	0,07	1,00	0,00
Satisfaction with the quality of university services	9	38,10	0,00	0,90	0,20	0,00
Satisfaction with extracurricular activity	5	37,89	0,00	1,00	0,00	0,00
Satisfaction with information support of curricular and extracurricular processes	6	39,05	0,00	0,00	1,00	0,00
Satisfaction with studying at university in general	11	38,85	0,00	0,15	1,00	0,00

Table 6: Educational process risk assessment.

Identified risks	Fuzzy characteristics				Risk probability
	N	AN	H	Cr	
1 Risk of knowledge obsolence	0,00	0,14	0,00	0,00	0,30
2 Mismatching of the stakeholders interests	0,00	0,14	0,00	0,00	0,30
3 Technical system malfunctioning	0,00	0,64	0,20	0,00	0,37
4 Risk of the key personnel dependence	0,25	0,45	0,25	0,25	0,43
5 Personnel depletion	0,00	0,14	0,17	0,00	0,46
6 Stagnation of research	0,00	0,07	0,00	0,00	0,30
7 Devaluation of personnel creativity	0,00	0,33	0,00	0,00	0,30
8 Lack of identity and uniqueness	0,37	0,75	0,37	0,37	0,41

5 CONCLUSIONS

The new idea of organizing social monitoring of the quality of university educational process as the assessment by the key players of educational relations – students and teachers has been discussed in the article. The method of representation of assessment grades as the university risk background characteristics by means of fuzzy composition allows to calculate probability measures for educational process risks.

The pilot testing of the described approach was done on the basis of processing data of the pilot sociological research through the survey of students and teachers of Siberian State University of Telecommunications and Information Science. The received assessment grades of satisfaction with different aspects of the educational process demonstrated at the moment of the survey its mean level, and transformation of the grades by means of risk thermometer into temperature indicator showed fever and risk background on the 'satisfaction' segment of data, which requires taking management decisions. It should be emphasized that random combinations of the grades demonstrate high degree of uniformity (variation coefficient does not exceed 3% in average), which proves the validity of the received results.

REFERENCES

[1] M. Bratti, A. McKnight, R. Naylor, and J. Smith, "Higher education outcomes, graduate employment and university performance indicators." in Journal of Royal Statistical Society A,167, part 3, pp. 475-496, 2004.

[2] V. Scherman and R.J.Bosker, "The Role of Monitoring in Enhancing the Quality of Education," in: Scherman V., Bosker R.J., Howie S.J. (eds) Monitoring the Quality of Education in Schools. SensePublishers, Rotterdam [Online]. Available: https://doi.org/10.1007/978-94-6300-453-4_1, 2017.

[3] I. V. Mitrofanova, "Multilevel monitoring as as social technology for managing the quality of professional education in modern Russia" cand. of sci.dissertation, Moscow, p. 158, 2009.

[4] M. A. Burova, "Social monitoring as a means of managing comprehensive education", Saratov, p.177, 2009.

[5] B. Williamson, "Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education," in British Journal of Educational Technology, vol. 50. N 6, pp. 2794-2809, 2019. doi:10.1111/bjet.12849.

[6] H. Lucas, M. Greely, and K. Roelen, "Real Time Monitoring for the Most Vulnerable: Concepts and Methods," in IDS Bulletin, vol. 44, N 2, pp. 15-30, March 2013.

[7] T.-E. Chen, et al., "An effective monitoring framework and user interface design," in Software: practice and experience, vol. 45, pp.549-570, 2015.

[8] Commonwealth Risk Management Policy [Online]. Available: https://www.finance.gov.au/sites/default/files/2019-11/commonwealth-risk-management-policy_0.pdf.

[9] J. Birkinshaw and H. Jenkins, "Making better risk management decisions," in Business strategy review vol.4, pp. 41-45, 2010.

[10] M. Adib and X.-Z. Zhang, "The risk-based management control system: A stakeholders' perspective to design management control systems," in International Journal of Management and Enterprise Development, vol.18, issue 1-2, pp. 20-40, 2019.

[11] J. Young, "The use of key risk indicators by banks as an operational risk management tool: a south African perspective," in Corporate Ownership & Control, vol. 9, issue 3, pp.172-185, 2012.

[12] A. A. Beloglazov, L. B. Beloglazova, O. V. Bondareva, and H. E.Ismailova, "Monitoring of the efficiency of teaching under conditions of education modernization and computerization," in Bulletin of the Russian Peoples' Friendship University, vol.14, N 2, pp. 220-232, 2017.

[13] E. Razinkina, et al., "Student satisfaction as an element of education quality monitoring in innovative higher education institution" in E3S Web of Conferences, vol.33 [Online]. Available: https://doi.org/10.1051/e3sconf/20183303043, 2018.

[14] Methodology of the procedure of accreditation expertise for university curricula, Moscow : Federal State Budget Organization "Rosaccredagentstvo", p. 164 [Online]. Available: http://nica.ru, 2015.

[15] T. S. Iljina, A. I. Baranova, and V. S. Kanev, "Management of the educational competence risk in tertiary education" in SibSUTIS Bulletin, vol. 1, pp. 3-11, 2017.

[16] O. M. Lopez, S. M. Hurtado, O. Botero, and F. Legendre, "Risk assessment methodology: Implementation of duration gap in corporate portfolios in order to reduce the systemic risk" in Estudios Gerenciales, vol. 34, issue 146, pp. 34-41, 2018.

[17] T. I. Monastyrskays, E. E. Gorjachenko, and N. L. Mikidenko, "Development and testing of the methodology of social monitoring for assessing the quality of education in SibSUTIS," report, Novosibirsk, p. 367, 2015.

[18] R. N. Ismailova, O. V. Krjukova, N. G. Nikolaeva, and E. V. Rakov, "Monitoring of consumer satisfaction" [Online]. Available: https://cyberleninka.ru/article/n/monitoring-udovletvorennosti-potrebiteley/viewer, 2014.

[19] V. N. Vjatkin and V. A. Gamza, "Risk-managemnet of a firm: the program of integrative risk-management," Moscow: 'Financy i Statistica' publisher, p. 400, 2006.

[20] A. Pegat, "Fuzzy modelling and management", 3-d edition, BINOM publisher, p. 801, 2015.

[21] M. Jevšček, "Competencies assessment using fuzzy logic" in Journal of Universal Excellence, vol. 5, pp. 187-202, 2016.

[22] M. Laal, "Knowledge management in higher education" in Procedia Computer Science, vol. 3, pp. 544-549, 2011.

[23] A. Varghese, Sh. Kolamban, S. Nayaki, and S. J. Prasad, "Outcome based Assessment using Fuzzy Logic" in International Journal of Advanced Computer Science and Applications, vol. 8, issue 1, pp. 103-106, 2017.

[24] Y. Shevtsova, V. Kanev, A. Poletaikin, and N. Kuleshova, "Optimizing Risk-Free Model of Development of Educational Organization Based on Modified Risk Thermometer" in materials of the 15th International Asian School-Seminar Optimization Problems of Complex Systems, Novosibirsk, Russia, pp. 68-72, 2019.

[25] N. V. Katilova and S. Angel, "The practice of key indicators for operational risks," in Financial risk management, vol. 2, pp. 190-204, 2006.

[26] X. Shi, Y.D. Wong, M.Z.F. Li, and C. Chai, "Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory" in Accident analysis and prevention, vol.117, pp.346-356, 2018.

A Review and Comparison of Mapping and Trajectory Selection Algorithms

Dmitrii Vershinin[1,2] and Leonid Mylnikov[1]
[1]*Perm National Research Polytechnic University, Komsomolsky avenue 29, 614990 Perm, Russian Federation*
[2]*Hof University of Applied Sciences, Alfons-Goppel-Platz-1, 95028 Hof, Germany*
nsenz@yandex.ru, leonid.mylnikov@pstu.ru

Keywords: SLAM, Scene Map, Trajectory, Cybernetics, Navigation, Autonomous Robot, Sensors.

Abstract: The dire need to solve orientation and localization tasks is directly related to the development of autonomous robotics systems as autonomous modules. In this article, we have reviewed and analyzed possible areas and peculiarities when implementing existing localization approaches in autonomous robotics systems operating under various weather conditions with possible obstacles on their way without preliminarily generated maps. In the paper, we especially pay attention to existing SLAM algorithms and a multitude of hardware concerned with this problem. Every considered and addressed algorithm in this paper comes with its main principles and generated, as a result of its performance, map type. The comparison of the algorithms was mainly based on the data of several articles and projects, in which almost perfect indoor experiments without any weather impact in order to examine the efficiency of the algorithms were conducted. Using the results acquired by the authors, a comparative table with main statistics for every considered algorithm was created. Apart from that, similar statistics for trajectory selection algorithms that meant to help researchers solve scenario/scripted tasks were covered. As a result of our review piece, we presented a ranging technique for the pair algorithms/sensors that uses the renowned TOPSIS outranking methodology. The proposed approach may become of significant help while selecting the pair for every case study.

1 INTRODUCTION

The rise of robotics dates back to the production systems efficiency problems – as a tool for assembling vehicles (cars in particular) and other complex units at factories. Today, similar robotics systems are frequently used in every aspect of human life. For example, there is a wide variety of both industrial and private cleaning robots that are also capable of scrubbing swimming pools and other surfaces, service or assistant robots and plenty of consulting robots at malls and airports. Besides, they are particularly good at helping handle with the aftermaths of technological accidents and meteorological disasters by getting to the hard-to-reach sites and seeking for casualties. Moreover, now they have become extremely popular amongst military services. For example, they can be used to collect intelligence and disarm bombs. Another fascinating use case for such robots is space in general and space exploration in particular. Not only are they used for repairing and fixing satellites, aero crafts and so on, but they also have the Moon and Mars exploration objective. Today they found their place in healthcare. Virtually they even help in agriculture and forestry. However, in order for them to be fully automotive and standalone, they need to have inside recognition, navigation and mapping algorithms. This set of algorithms and problems is commonly referred to as SLAM (Simultaneous Localization and Mapping).

These days there are too many approaches and methods for solving SLAM problems, however, only recently we have been able to notice a distinct transition to modern (hybrid) techniques and approaches, that are capable of processing data with outliers without any pre-processing, from the conventional ones (i.e. filter-based algorithms).

Additionally, in mobile robotics there is a substantial problem with processing every localization and mapping task simultaneously (S letter in SLAM) and concurrently.

Even more complex problem is when we are supposed to solve the outlined tasks in an unknown or/and unstructured and dynamically changing environment with many obstacles whether there are bump/altitude variations or sudden climate changes

affecting the precision and accuracy of the sensor data.

Furthermore, depending on the algorithm, an operating robot may consider using pairs of sensors of different types: acoustic, lidars, cameras, sonars, altimeters. Likewise, it is crucial to consider possible precision and accuracy requirement when opting for one or the other algorithm or sensor, since, for example, lidars are fairly efficient yet exceedingly expensive.

The type of the environment encompassing our robotics system should be considered as well, as in self-resembling and similar scenes (i.e in office rooms) the precision and accuracy requirements are likely to become stricter due to lots of outliers concerning false-positive loop closures (making the robot think that it has already visited this place).

Thus, a particular algorithm that one wants to use will be affected by the collection of sensors we have at our disposal or can use, which may constitute another worth considering and requiring a comprehensive analysis problem. And given the fact that most modern approaches are hybrid-based (it is possible to use different combinations of sensors and algorithms at run time), the mentioned selection task becomes crucial since under distinct conditions (weather, traffic, obstacles related) the efficiency of the algorithms/sensors pair may vary drastically, which is likely to lead to unpredictable results. For instance, in [1] the influence of changing weather conditions on the SLAM results of automated vehicles was shown, and in [2] authors showed how illumination and sensors positioning affect the quality of the results (i.e. the target may be behind vegetation). In [3] an autonomous adaptive multisensor SLAM was demonstrated.

2 A REVIEW OF SLAM ALGORITHMS

When comparing SLAM algorithms, the most fundamental aspect is the resulting trajectory of these algorithms. In this context let us consider trajectory as a set of points in space that are dependent on their coordinates, ambient influence on it associated with the noise of sensor data, obstacles and the positional changes of dynamic objects and obstacles. Thereby, trajectory and its possible changes are reliant on the dimensions of a robotics system. In general, a trajectory may be illustrated as a chain of interconnected coordinates, which can be seen in Figure 1.

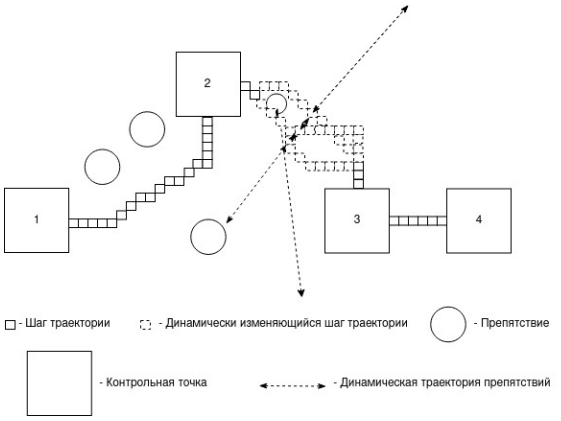


Figure 1: A trajectory example.

Accordingly, a trajectory step will be represented as follows: (x_i, y_i, t_i), where x_i- a robot's x position on the map, y_i- a robot's y position on the map, t_i- coordinates registration time. When there are errors/environmental influence present, the same trajectory will take the following form: $((x_i \pm c_i), (y_i \pm d_i), (t_i \pm \Delta))$, where c - external influence on the x position, d - external influence on the y position, Δ - registration deviations.

Apart from that, every SLAM algorithm generates a map of a robot's environment. This way let us represent a map as a manifold of points on the space grid containing their coordinates and probabilities of obstacles located in these areas (including the dynamic ones). But for 3D cases - a cube of grids populated with points in space. The scheme of the map is represented in Figure 2.

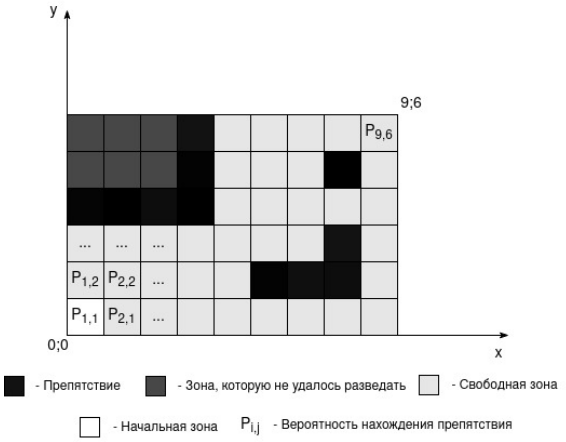


Figure 2: A schematic representation of an environmental map.

In this case P – the probability indicating a present obstacle in a particular cell of the grid. The cells that have been visited by the robot are coloured light-grey,

dark-grey – for uncharted cells and black - for walls or static obstacles.

2.1 SLAM Algorithms

Hector SLAM – is a SLAM method that operates by means of extracting data from a 2D lidar. At the moment, it is one of the most popular approaches that is, on top of that, widely used in a variety of mobile robotics projects. The algorithm builds a 2D map and provides localization possibilities at the scan rate of the lidar. In order to build a correct map, a conversion from the local lidar‘s coordinate system to the surface, that the robot is moving through, coordinate system should be performed [4].

Hector SLAM builds an occupancy grid (the map that corresponds to the one described in our definition), in which every cell is coloured: black – the cell is occupied, light-grey – the cell is empty, dark-grey – the cell has not been checked yet. An example of the resulting map can be seen in Figure 3.

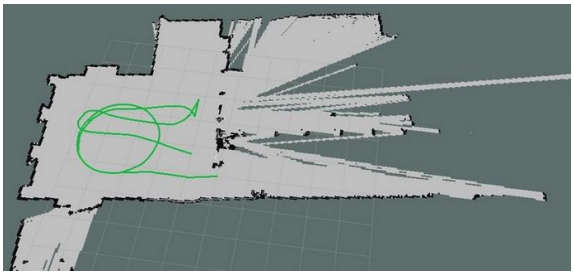


Figure 3: A Hector SLAM resulting map [5].

ORB-SLAM – is a versatile and accurate SLAM algorithm, based on features recognition and real-time trajectory calculation utilizing monocular cameras, which builds an environment sparse 3D scene map. It can close large loops and perform global relocalisation in real-time and from wide baselines. Apart from that, ORB-SLAM makes it possible to automatically initialize scenes of different types [6]. The resulting map of this algorithm is a sparse 3D map, an example of which is illustrated in Figure 4.

Figure 4: An ORB-SLAM resulting map [7].

DPPTAM – is one of the newest visual SLAM algorithms, which adjusted and implemented the most successful ideas of the previous algorithms. It is a direct monocular odometry algorithm that estimates a dense reconstruction of a scene in real-time on CPU and saves the trajectory as a sequence of points in the particles cloud. To build high-resolution images the algorithm makes use of standard techniques for minimizing the points errors [8]. An example of this map is shown in Figure 5.

Figure 5: A DPPTAM resulting map [9].

ZEDfu – tracks positioning and orientation based on a ZED camera mounted on a tracking device. A ZED camera builds a real-time 3D world and recognizes rooms and objects. As a resulting map, the algorithm builds a 3D lattice from particles clouds of any environment (either indoors or outdoors) [10]. An example of this map can be seen in Figure 6.

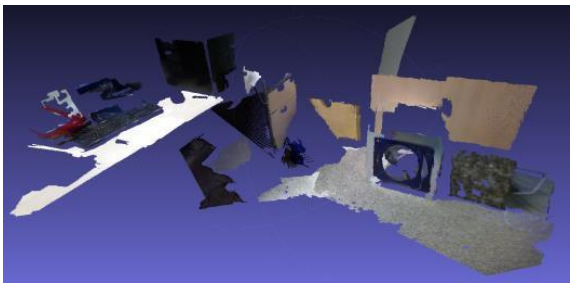


Figure 6: A ZEDfu resulting map [10].

RTAB-Map – is a RGB-D, Stereo and Lidar Graph-Based SLAM approach based on an incremental appearance-based loop closure detector. The loop closure detector uses a bag-of-words approach to determinate how likely it is that a new image comes from a previous location or a new location. When a loop closure hypothesis is accepted, a new constraint is added to the map's graph, then a graph optimizer minimizes the errors in the map. RTAB-Map can be used either with a Kinect or with a stereo camera/lidar [11]. An example of this map is shown in Figure 7.

Figure 7: An RTAB-Map resulting map [11].

Using such SLAM methods, we can build an environmental map and localize the considered object in space, however, in order to solve possible scenario tasks, perform various operations (i.e. holding or raising objects, approaching them in the most efficient way and performing manipulations with them) we may need to use proper trajectory selection algorithms.

An overall comparison of the algorithms' characteristics can be seen in Table 1.

2.2 Trajectory Multi-Objective Optimization Algorithms

MOACO is a multi-group trajectory optimization algorithm. MOCAO uses multiple pheromone matrices and more heuristic matrices. Each of these matrices is responsible for only one task. All the agents are divided into several groups. Each group has multiple weights and each agent in the group has its weight vector. If the number of weights in the group is less than the number of agents, then the other agents are set weights from the beginning. Thus, two or more agents in a group may use the same weight vector. But every agent uses its vector to aggregate the pheromone and heuristic information. Afterwards, it calculates its probability to move to an unvisited spot and chooses the next spot to visit via wheel roulette selection [13]. Finally, it uses non-dominated solution generated by current iteration to update the pheromone information.

MACS uses one pheromone matrix and multiple heuristic matrices. Each heuristic matrix is responsible for only one task. Each agent has a weight vector and all the heuristic matrices are aggregated by weighted product. The weight vectors between two distinct agents are different and non-dominant solution is used to update the pheromone information. MACS considered to be a similar to MOAQ approach with only one difference in the number of the weight vectors used to aggregate heuristic information: in MOAQ two such vectors are used {0,1} and {1,0}, whereas MACS uses more vectors owing to its influence on the algorithm [13].

PACO algorithm uses multiple pheromone matrices and only one heuristic matrix. All the agents share this single matrix. Each pheromone matrix is responsible for only one task. As in MACS algorithm, each agent has its weight vector and all the pheromone matrices are aggregated by the weighted sum. This algorithm uses the best option and the second to the best solution of each objective to update pheromone information. The non-dominated solutions mainly approximate to the central part of the Pareto front [13].

MOEA is one of the most efficient agent algorithms. The algorithm performs multi-object decompositions and simultaneously optimizes its parts. Each subtask is optimized by using the nearby subtasks information [13].

The algorithms comparison is presented in Table 2.

Table 1: SLAM Algorithms comparison (based on [12] data).

Algorithm	Accuracy	Algorithm type	Quality of the map	Best suited for	Odometry quality	Max deviation (м)	Obstacles avoidance precision	RMSE(м)	Mean (м)	Std (м)
Hector	High	Occupancy grid	Good	2D lidar	–	0.18	Excellent	0.088	0.025	0.024
ORB	High	Feature based map	Low	Monocular camera	Good (0.43м)	0.43	Good	0.166	0.159	0.047
DPPTAM	Average	Cloud of particles based map	Average	Monocular camera	Bad (4.26м)	4.26	Good	0.338	0.268	0.206
ZEDfu	Very high	Cloud of particles based map	Good	Stereo ZED camera/Kinect	Good (0.32м)	0.32	Good	0.726	0.631	0.358
RTAB	Above average	Cloud of particles based map	Good	Kinect	Good (0.67м)	0.67	Average	0.163	0.138	0.085

Table 2: Trajectory multi-objective optimization algorithms comparison (based on [13] data).

Algorithm	Max (м)	Min (м)	Mean (м)	Std (м)
MOACO (kroAB100)	0.3412	0.3037	0.3236	0.0105447
MACS (kroAB100)	0.1924	0.1675	0.1823	0.0054743
PACO (kroAB100)	0.4076	0.3695	0.3912	0.011176
MOEA (kroAB100)	0.1062	0.0505	0.0767	0.0144815
MOACO (kroAC100)	0.3475	0.3211	0.3352	0.0063372
MACS (kroAC100)	0.1995	0.1723	0.1885	0.0051235
PACO (kroAC100)	0.2612	0.2285	0.2413	0.0086281

3 HARDWARE USED BY THE ALGORITHMS

As it has been mentioned before, the selection of a sensor for a particular algorithm is an essential task. It is mainly due to the technical characteristics of the sensors and their use case limitations. Thus in this section, an excerpt from a paper on sensors efficiency when a variety of obstacles is present will be presented.

First of all, the most conventional robotics sensor types should be enumerated and described here:

- Proximity sensors – detect objects that are located in close proximity to the robot. These sensors can detect objects' presence by using light, sound or electromagnetic fields (for example, infrared, ultrasound sensors and LDRs).
- Rangefinders – determine the distance between two distinct objects in an environment (for example, cameras, lasers, lidars).
- Tactile sensors – provide information about physical contacts with objects.
- Light sensors – detect light density that consequently can be converted into current or voltage.
- Sound sensors – detect sound and return proportional to the sound level voltage.
- V/I Converters.

Let us consider the most common sensors for detecting objects and obstacles.

RADAR/LIDAR. To detect obstacles the RADAR (radio detection and ranging) /LIDAR (light detection and ranging) combination is frequently used.

Detection and distance measurements are one of the main LIDAR functions. The distance is represented as the time required for a light impulse to travel from a sender to a photodetector after its reflection from an object/obstacle surface. The distance is defined as: , where – represents the distance, – the speed of light, – the impulse time. Therefore, LIDAR can obtain objects' 3D geometry [12] and [14].

Cameras are probably the most popular sensors used to detect objects and environmental changes.

Their main peculiarity is in the ability to recreate a 3D cloud of particles of a particular environment.

Table 3: Sensors' characteristics comparison (based on [14] data).

Sensor's type	Size	Power consumption (Вт)	Depth (м)	Price ($)	Effective precision	Best suited for	Acceptable temperature range	Requires additional hardware	Dependent on illumination	How representative for people	Efficiency under bad weather conditions
Acoustic	Very compact	0.01 - 1	2-5	10-500	1-3% of max depth	Range/acoustic based SLAM	From -20 Up to +80	-	-	Bad	Weather resistant but operate poorly in noisy environments
Monocular cameras	Very compact	0.01 - 10	-	100-5k	25–135 mm	Visual SLAM	Depends on a camera, however in general these cameras have a pretty decent temperature resistance	+	+	Good	Weather resistant provided the lenses are clean
Lidar	Bulky	50 - 200	50-300	5k-100k	Up to ± 3 cm	Range/distance based SLAM	From -50 Up to +80	-	-	Bad	Efficiency drops in rainy, snowy and foggy conditions
Stereo cameras	Compact	2 - 15	5-20	500-5k	1mm - 5см	Visual SLAM	-	+	+	Good	Weather resistant provided the lenses are clean

Apart from that, cameras are able to detect obstacles by creating depth maps from consecutive images fetched by monocular cameras. However, this approach works perfectly with only static obstacles [12], [14].

SONAR (Sound navigation and ranging) – creates a sound impulse and measures the impulse's echo return time. Therefore, the sensor's results cannot be affected by light or illumination. However, sonars are mainly used for nearby detections, meaning that they are futile when it comes to measuring distant objects. But, an even worse drawback of the type is in its inability to operate in noisy environments (engine vibrations, highways, toots) as was shown in [12] and [14].

Laser rangefinder. The measurement principle is based on the angle between the laser ray pointing at an object and the laser's lens. Having this laser-lens distance (h) and the angle, we can calculate the distance to the object – the less the angle, the farther the object [12] and [14].

The comparison of the sensors' characteristics is represented in Table 3.

The sensors that can operate under well-illuminated conditions may easily avoid smoke conditions or even mist. However, when it comes to cameras, it is important to make them able to use infrared or thermal vision in order to increase their performance under such conditions.

Under rainy conditions, the most efficient sensors are LIDARs, Laser rangefinders and some types of cameras.

When operating in a blizzard, the most efficient sensors are cameras, since LIDAR/RADAR systems may be covered with snow, which prevents them from delivering any acceptable result.

Working in a highly reflective environment, neither of the sensors without supportive filtering algorithms demonstrated their efficiency.

When there are physical obstacles present (slopes, hills, slides etc.) the most efficient systems are LIDAR and multi-camera systems [1].

Owing to combinations of monocular cameras, it becomes possible to detect a wide range of obstacles around our object. However, in comparison with the cameras, LIDARs provide much better precision and FOV.

However, sometimes LIDAR data becomes incorrect due-to its distance to an object/obstacle. As in the cameras' case, some of the LIDAR sensors may be exceedingly noise sensitive.

Therefore, in a wide variety of modern robotics systems, researches use hybrid approaches. For example, some of them propose a combination of 6 SONARs with 3 Visual cameras for obstacles

detection under any condition. Besides, it is a good practice to use LIDARs and RADARs together to reduce the resulting errors. In order to boost the performance even further, it is helpful to add visual cameras to the aforementioned combination to be able to obtain information about roads, road signs, signals, etc [1].

So, we can conclude that these hybrid systems, which use a set of SLAM algorithms, are becoming more and more popular and relevant due to the importance of changing of adapting SLAM algorithms at runtime.

4 RESULTS

As a result of the review, we can formulate a trajectory/SLAM selection algorithm that will make it easier for researches to opt for a particular set of sensors/algorithms and take into account existing restrictions and constraints (see Figure 8).

Step 1: Define the robotics systems restrictions and constraints (financial, technical) and the system requirements (ability to operate under different weather conditions, avoid obstacles, operate in noisy environments etc.).
Step 2: Formulate a list of parameters/measurements from the restrictions and requirements.
Step 3: Perform the ranking operation of the solutions (individually for algorithms and hardware) by pairwise considerations of the parameters/measurements as shown in Figure 9.
Step 4: Based on the ranging results the expert should choose a pair(s) of the algorithm(s)/sensor(s) depending on their priorities and preferences. The initial selection of several pairs will sift the range of available options making it easier for them to expertly select the most suitable option.

Figure 8: The selection algorithm for choosing the most suitable methods and sensors for robotics systems operating in any environment.

To perform the third step one may make use of some of the following outranking methods: TOPSIS [15], ELECTRE, VIKOR, PROMETHE. For example, in Figure 9 the ranking criteria are represented as x and y axes, points represent the selected algorithms from the previous steps. To include a researcher's subjective point of view, a particular sign is plotted on this chart (in this case, it is a diamond). Then, the same ranking approach should be performed for the remaining parameters/measurements. Based on the distance from the best and the worst option the researcher will be able to select the most suitable algorithms.

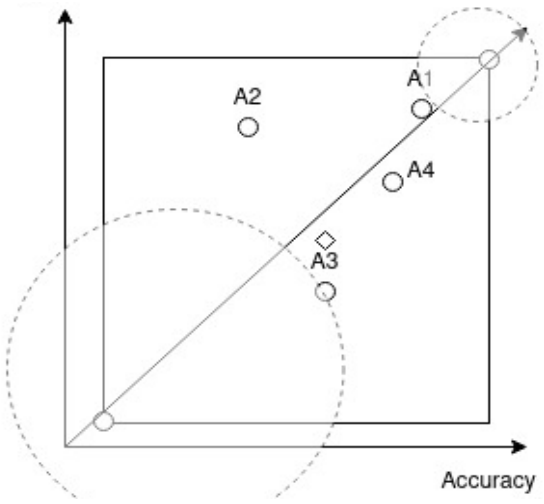


Figure 9: A ranking example (A1 - Hector SLAM, A2 - DPPTAM SLAM, A3 - ORB SLAM, A4 - RTAB SLAM, prices were set as approximate based on the previous tables).

5 CONCLUSIONS

There are no comprehensive robotics systems that are able to operate under any potential condition. Therefore, based on a detailed analysis (like this one), we can select the most appropriate set of algorithms/sensors for our particular case to operate under required conditions (weather, obstacles related). However, it is crucially important to understand that apart from the covered conditions, there might be some additional ones (for example, related to financing the project, difficulty of the project, personal preferences, the experience of the team and so on) and, in this case, the eventual choices may vary. Therefore, this paper cannot solve or cover all the problems regarding that selection, but this review may still be considered as a useful handbook.

According to this article we may conclude that at the moment there are two approaches to solving SLAM problems regarding the algorithmic part: 1) Based on our requirements, we can create new algorithms and systems by combining the existing ones and 2) Develop brand new algorithms that would entirely solve our problems based on the requirements and limitations.

REFERENCES

[1] X. Yu and M. Marinov, "A study on recent developments and issues with obstacle detection systems for automated vehicles," Sustain., vol. 12, no. 8, 2020, doi: 10.3390/SU12083281.

[2] P. Slivitsin, A. Bachurin, and L. Mylnikov, "Robotic system position control algorithm based on target object recognition," Proc. Int. Conf. Appl. Innov. It, vol. 8, no. 1, pp. 87-94, 2020.

[3] Sh. Shen, "Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft micro-aerial vehicle (MAV)," US10732647B2, 2013.

[4] S. Kohlbrecher and J. Meyer, "Hector SLAM" [Online]. Available: http://wiki.ros.org/hector_slam, 2012.

[5] W. A. S. Norzam, H. F. Hawari, and K. Kamarudin, "Analysis of Mobile Robot Indoor Mapping using GMapping Based SLAM with Different Parameter," IOP Conf. Ser. Mater. Sci. Eng., vol. 705, no. 1, 2019, doi: 10.1088/1757-899X/705/1/012037.

[6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Trans. Robot., vol. 31, no. 5, pp. 1147-1163, 2015, doi: 10.1109/TRO.2015. 2463671.

[7] M. Sokolov, O. Bulichev, and I. Afanasyev, "Analysis of ROS-based visual and lidar odometry for a teleoperated crawler-type robot in indoor environment," ICINCO 2017 - Proc. 14th Int. Conf. Informatics Control. Autom. Robot., vol. 2, no. July, pp. 316-321, 2017, doi: 10.5220/0006420603160321.

[8] A. Concha and J. Civera, "DPPTAM: Dense piecewise planar tracking and mapping from a monocular sequence," IEEE Int. Conf. Intell. Robot. Syst., vol. 2015-December, no. July, pp. 5686-5693, 2015, doi: 10.1109/IROS.2015.7354184.

[9] StereoLABS, "ZEDfu" [Online]. Available: https://www.stereolabs.com/docs/.

[10] I. Z. Ibragimov and I. M. Afanasyev, "Comparison of ROS-based visual SLAM methods in homogeneous indoor environment," 2017 14th Work. Positioning, Navig. Commun. WPNC 2017, vol. 2018-January, no. October, pp. 1-6, 2018, doi: 10.1109/WPNC.2017. 8250081.

[11] M. Labbé and F. ¸Michaud, "RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation," J. F. Robot., vol. 36, pp. 416-446, 2019.

[12] M. Filipenko and I. Afanasyev, "Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment," 9th Int. Conf. Intell. Syst. 2018 Theory, Res. Innov. Appl. IS 2018 - Proc., no. November, pp. 400-407, 2018, doi: 10.1109/IS.2018.8710464.

[13] J. Ning, C. Zhang, P. Sun, and Y. Feng, "Comparative study of ant colony algorithms for multi-objective optimization," Inf., vol. 10, no. 1, pp. 1-19, 2018, doi: 10.3390/info10010011.

[14] M. Zaffar, S. Ehsan, R. Stolkin, and K. M. D. Maier, "Sensors, SLAM and Long-term Autonomy: A Review," 2018 NASA/ESA Conf. Adapt. Hardw. Syst. AHS 2018, pp. 285-290, 2018, doi: 10.1109/AHS.2018.8541483.

[15] Z. Pavić and V. Novoselac, "Notes on TOPSIS Method," Int. J. Res. Eng. Sci., vol. 1, no. 2 [Online]. Available: https://www.researchgate.net/ publication/ 285886027_Notes_on_TOPSIS_Metho%0Awww.ijres.org, pp. 5-12, 2013.

Development of Physical and Psychological States Graphs of People and Their Software Implementation in the Tasks of Evacuation Modelling

Ekaterina Yurchenko[1], Irina Shulga[1], Mikhail Tugarinov[1], Igor Shelekhov[2] and Stanislav Torgaev[1]
[1]*Faculty of Radiophysics, Tomsk State University, Lenin avenue 36, 634034 Tomsk, Russian Federation*
[2]*Faculty of Psychology and Special Education, Tomsk State Pedagogical University, Kievskaya Str. 60, 634061 Tomsk, Russian Federation*
kattifi@mail.ru, shulga.irina20762@yandex.ru, mtugarinov@mail.ru, brief@sibmail.com, torgaev@mail.tsu.ru

Keywords: Evacuation, Fire, Graph, Physical State, Behavior, 3D Modelling, Emergency, Block Programming.

Abstract: The purpose of the presented in the article results is to increase the realism of the people evacuation modeling in case of emergency situation. Models that exist today do not describe in detail physical and psychological states of the characters during the simulation. This article presents the results of the development of people physical and psychological states graphs in the conditions of evacuation. All graphs are presented as extended final state machines. On the basis of the developed finite state machines the description of state transitions was carried out and algorithms were built This work was carried out as part of the development of a comprehensive 3D model of the people evacuation processes of in emergency situations in particular fires. A software implementation in the *Unreal Engine* program of these states was performed. Examples of the behavior of characters in various psychological and physical states are also presented.

1 INTRODUCTION

To date, methods of modeling various processes are actively developing. Modern modeling methods and tools allow us to study both technical and social processes.

In particular, there is a large number of works aimed at modeling the processes of evacuation of people in emergency situations [1-12]. Depending on the approach there are 4 types of models: the molecular approach [1,2], the route-based approach [3,4], the group-based approach [5-7] and the agent-based approach [8-10]. The most promising models are those built using an agent-based approach.

One of the main problem of evacuation models developing is making the simulation more realistic. The most realistic modelling requires detailed 3D models with a rich set of simulated parameter. In developing such models, special attention should be paid to modelling character's behaviour. Existing 2D and 3D models have a large number of simplifications, especially in terms of the psycho-physical state of characters [1-12].

In this regard, the purpose of this study is to develop graphs describing the physical and psychological states of people in emergency situations, in particular, in case of fire. This task requires taking into account various physical parameters of the characters, as well as detailed modeling of their psychological state and behavior. In our work, the description of states and transitions between them is performed in the form of an extended finite state machine. Automatic description will further simplify the process of analysis, testing and verification of the model through the use of tools of the automata theory [13-14].

The development of such graphs (extended finite state machines) will allow to obtain algorithms for transitions between states on the basis of which their software implementation will be performed. The results of the development of state graphs and software implementations of the transition algorithms will form the basis of the 3D modeling system of evacuation processes with an increased degree of realism of the model characters behavior. The implementation of these algorithms and the 3D modelling system based on them is carried out in the

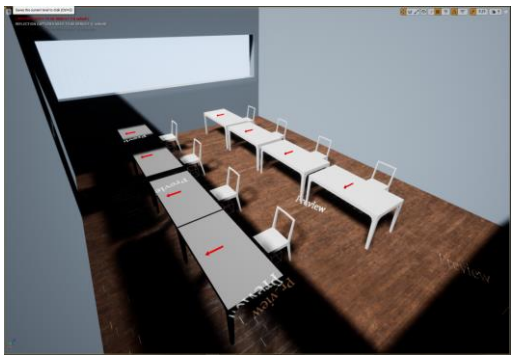
Unreal Engine environment. This environment has great functionality and in the future will allow the addition of VR technology into the model.

2 MODEL DESCRIPTION

As noted above, the development of a detailed 3D model is carried out in the *Unreal Engine* environment. The programming of the characters behavior was made using the *Blueprint* visual programming tool. In this tool all processes are represented as blocks that contain a specific code that performs the required action. There are many types of such blocks: event blocks, action blocks and auxiliary blocks. The *Blueprint* visual programming tool allows to implement all possible actions of the character during evacuation.

To conduct simulation experiments in the *Unreal Engine* examples of test environments (buildings, rooms) were designed and algorithms for the fire and smoke propagation were implemented [12]. Examples of the test environments are shown in Figure 1.

a)



b)

Figure 1: Examples of test environments: a) building; b) room.

During the simulation each character has a set of parameters. Some of parameters are constant and do not change during the simulation. These parameters are: *weight, age, readiness for emergencies, temperament, leader* (the ability to lead people). The parameters to change are: *health*, *speed* and *stress*. In the simulation these parameters and processes of interaction with the environment will determine the physical state of each character and his behavior during the evacuation. The model considers four possible *Temperaments: Melancholic, Choleric, Sanguine and Phlegmatic*. This parameter determines the psychological state of the characters and their actions.

3 PHYSICAL AND PSYCHOLOGICAL STATES GRAPHS

To increase the realism of the simulation of the evacuation process we have developed extended finite state machines describing the physical and psychological states of the characters. These states of the character will determine its behavior during the simulation. The development of the extended final state machines was carried out in conjunction with a psychologist from Tomsk State Pedagogical University. The involvement of specialists in the field of psychology will significantly increase the realism of behavioral modeling.

3.1 The Physical States Graph

The physical states extended finite state machine is described as

$$M_{Physical\ state} = \{S, X, Y, V, T\},$$

where S is a finite set of states; X is a set of input symbols; Y is a set of output symbols; V is a set of context variables; T is a set of transitions between states [13-14].

The set of states S provides five physical states for each character is

$$S = \{S_1, S_2, S_3, S_4, S_5\},$$

where S_1 – *Initial state;* S_2 – *Intoxication;* S_3 – *Injury;* S_4 – *Intoxication/Injury;* S_5 – *Death.*

The input and output symbol sets are described as

$$X = \{smoke, fire, hit, health\},$$
$$Y = \{health, stress\}.$$

Input symbols are *smoke, fire, hit*: *smoke* is an impact of smoke on the character; *fire* is an impact of fire; *hit* is an external hit (construction or other character); *health* is a character's health parameter. The set of context variables is the same as the set of output symbols.

The set of state transitions is defined as

$$T_{n\to m} = \{S_n, x, P, up, S_m\},$$

where S_n – initial state; x – input parameters; P – predicate (transition condition); up – update function; S_m – final state.

Figure 2 shows an extended finite state machine of transitions from one physical state to another.

In the *Initial state* (S_1) the character has the maximum values of the *Health* and *Speed* parameters and minimum value of the *Stress* parameter. The transition to the state of *Intoxication* (S_2) occurs when the conditions for finding the character in the smoke at least some minimum time (Figure 3).

The transition to the state of *Injury* (S_3) occurs under two conditions: external impact (hit) and direct contact with fire for a minimum time of fire interaction (Figure 4).

If the character was in the *Intoxication* state and is exposed to fire or impact, then he goes into the *Intoxication/Injury* state. Similarly, the transition from the state of *Injury* is carried out when exposed to smoke (Figure 5).

$$T_{1\to 2} = \{\langle S_1, smoke, t > t_{\min}, (health--, stress++), S_2\rangle\}.$$

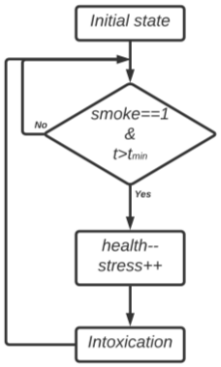


Figure 3: The $T_{1\to 2}$ transition algorithm to the state of *Intoxication*.

$$T_{1\to 3} = \{\langle S_1, fire, t > t_{min}, (health--, stress++), S_3\rangle,$$
$$\langle S_1, hit, , (health--, stress++), S_3\rangle\}.$$

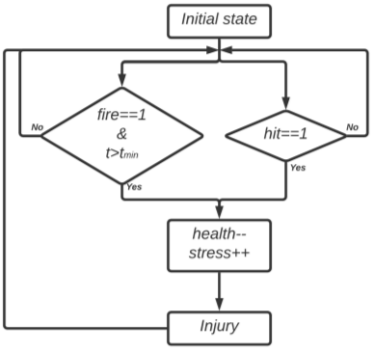


Figure 4: The transition algorithm to the *Injury state* (S_3).

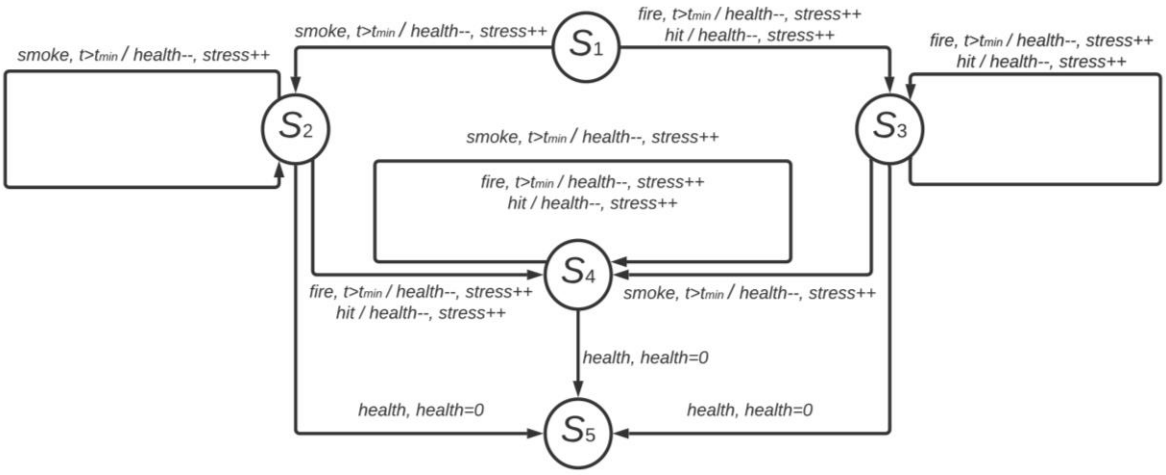


Figure 2: Physical states extended finite state machine.

$$T_{3\to4}=\{\langle S_3, smoke, t>t_{\min}, (health--, stress++), S_4\rangle\},$$
$$T_{2\to4}=\{\langle S_2, fire, t>t_{min}, (health--, stress++), S_2\rangle,$$
$$\langle S_2, hit, , (health--, stress++), S_4\rangle\}.$$

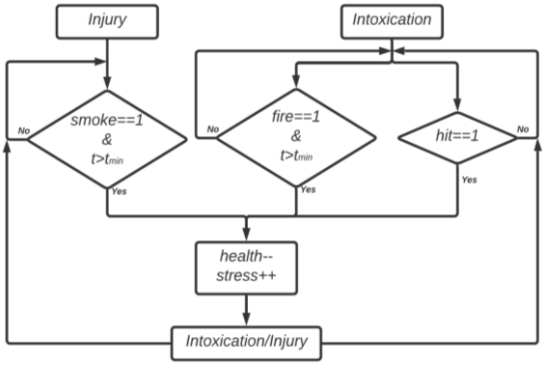


Figure 5: The transition algorithm to the state of *Intoxication/Injury* (S_4).

During the transition to *Injury*, *Intoxication* or *Intoxication/Injury* states there is a decrease in the *Health/Speed* and an increase in the *Stress* parameters of the character (Figure 2).

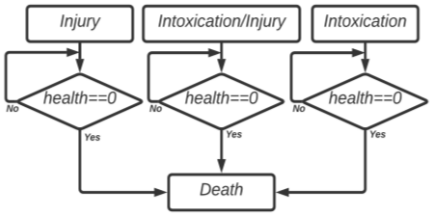


Figure 6: The transition algorithm to the state of *Death* (S_5) from S_2, S_3, S_4 states.

The transition to the *Death* state is possible from all states (except the *Initial state*) if *Health* parameter is equal to zero (Figure 6):

$$T_{2\to5}=\{\langle S_2, health, health=0, , S_5\rangle\};$$
$$T_{3\to5}=\{\langle S_3, health, health=0, , S_5\rangle\};$$
$$T_{4\to5}=\{\langle S_4, health, health=0, , S_5\rangle\}.$$

3.2 The Psychological State Graph

The psychological state graph is described similarly to the physical extended finite state machine:

$$M_{Psychological\ state}=\{S, X, Y, V, T\}.$$

The set of states S also provides five states for each character:

$$S=\{S_1, S_2, S_3, S_4, S_5\},$$

where $S_1 - Calm;\ S_2 - Panic;\ S_3 - Psychology;\ S_4 - Sympathy;\ S_5 - Group.$

The input and output symbol sets are described as

$$X=\{stress, leader, EmPr\},$$
$$Y=\{action\}.$$

Input symbols are *stress, leader, emergency preparedness (EmPr)*. The set of output symbols contains an *action* symbol which is defined in each specific state according to separate rules.

The set of transitions state is defined as in the previous case

$$T_{n\to m}=\{S_n, x, P, up, S_m\}.$$

The extended finite state machine of the psychological states of the characters is shown in Figure 7.

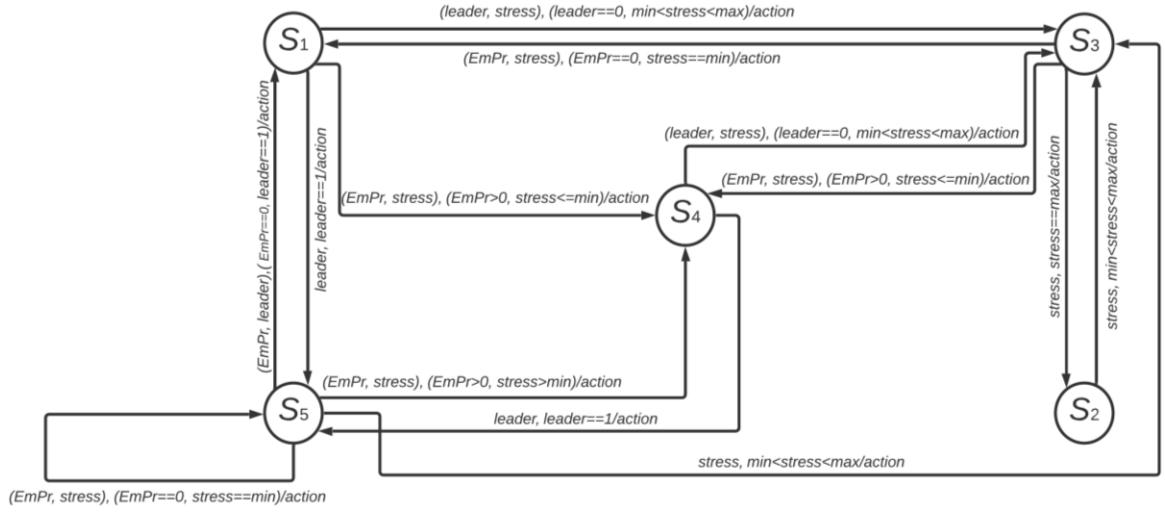


Figure 7: Psychological states extended finite state machine.

The *Psychology state* (S_3) have three sub-states: *Stupor*, *Aggression* and *Selfishness*.

The state of *Calm* corresponds to the appropriate behavior of the character. In this state the *stress* parameter has a minimum value. The transition to the *Psychology* state is performed if the value of the *stress* parameter becomes higher than the minimum. The transition to the state of *Sympathy* is performed if the *EmPr* parameter has a non-zero value and a character with a *health* parameter below the threshold value will fall into the field of view of this character. The transition to the *Group* state is performed if the group led by the *leader* falls into the field of view of the character.

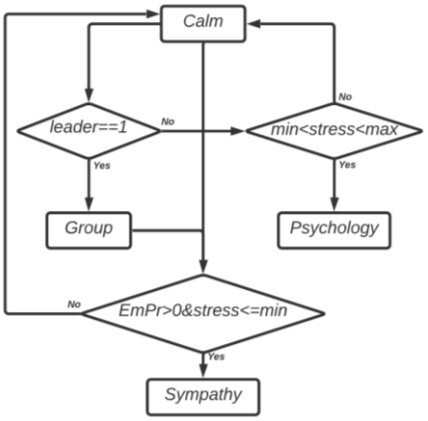


Figure 8: The algorithm of transitions from *Calm* state.

The transitions from *Calm* state can be described as (Figure 8):

$$T_{1\to3}=\{\langle S_1, leader, stress, leader=0 \,\&\, \max > stress > \min,\ , S_3\rangle\};$$

$$T_{1\to5}=\{\langle S_1, leader, leader=1,\ , S_5\rangle\};$$

$$T_{1\to4}=\{\langle S_1, stress, EmPr, Em.Pr.>0 \,\&\, stress <= min,\ , S_4\rangle\}.$$

The transition to the *Panic* state (S_2) occurs only from the *Psychology* state at the maximum value of the *stress* parameter of the character. In the *Panic* state, the character behaves inappropriately and makes chaotic movements. Each character in this state forms a "panic radius" around him. Some characters in this radius may increase the *stress* parameter. The transition back to the *Psychology* state is performed if the value of the *stress* parameter drops below the threshold value.

The transitions from *Panic* state can be described as (Figure 9).

As noted above the *Psychology* state (S_3) has three sub-states. In the *Stupor* substate the character is completely immobilized. Depending on the values of the *stress* parameter characters of all *Temperaments* can fall into this substate. In the *Aggression* substate the character can cause both physical and moral damage to others characterы. Only characters with a *Choleric* temperament can fall into this substate. In the *Selfishness* substate the character performs actions to rob the property of a certain territory and other characters. Only characters with a *Sanguine* temperament can fall into this substate. The transition to the state of *Sympathy* from the state of *Psychology* is carried out if the *stress* parameter drops to a value below the average, the *EmPr* parameter is not equal to 0 and a other character with the *health* parameter below the threshold value will come into view. And the transition to the *Calm* state is performed if the value of the *stress* parameter drops to a value below the average. The transitions from *Psychology* state can be described as (Figure 10).

$$T_{2\to3}=\{\langle S_2, stress, \max > stress > \min,\ , S_3\rangle\}.$$

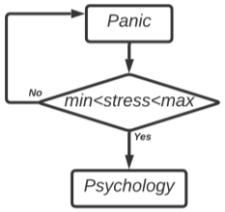


Figure 9: The algorithm of transitions from *Panic* state (S_2).

$$T_{3\to2}=\{\langle S_3, stress, stress=\max,\ , S_2\rangle\};$$

$$T_{3\to1}=\{\langle S_3, (EmPr, stress), Em.Pr.=0 \,\&\, stress=min,\ , S_1\rangle\};$$

$$T_{3\to4}=\{\langle S_3, EmPr, Em.Pr.>0,\ , S_4\rangle\}.$$

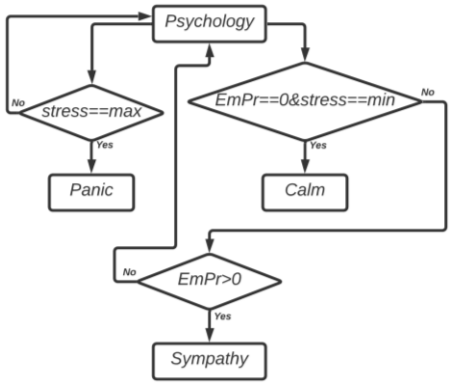


Figure 10: The algorithm of transitions from *Psychology* state (S_3).

The *Sympathy* (state S_4) is a state that represents a character prepared for emergencies. This character has the skills to help other characters with a

decreased *health* parameter and an increased *stress* parameter. The transition to the *Psychology* state is carried out if the value of the *stress* parameter becomes higher than the minimum. And the transition to the *Group* state is carried out when the organized group will fall into the field of view of this character. In this case the character joins the group. The transitions from *Sympathy* state can be described as (Figure 11):

$$T_{4\to5} = \{\langle S_4, Leader = 1, S_5\rangle\};$$
$$T_{4\to3} = \{\langle S_4, Leader = 0 \,\&\, \max > Stress > \min, S_3\rangle\}.$$

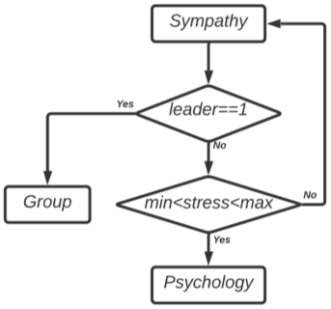


Figure 11: The algorithm of transitions from *Sympathy* state (S_4).

A *Group* (state S_5) is a state in which there is one character who is the *Leader*. The leader is the character with the maximum value of the *EmPr* parameter. The leader does not help the victims if they fall into his field of vision. If the *stress* parameter of the *Leader* increases to the average value or his *health* parameter decreases to the minimum, then he stops leading the group and the group breaks up. The speed of the group depends on the speed of the *Leader* (Figure 12).

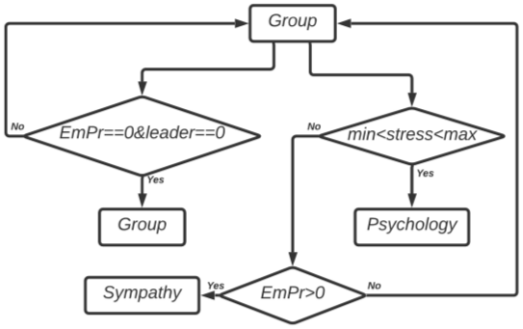


Figure 12: The algorithm of transitions from *Group* state (S_5).

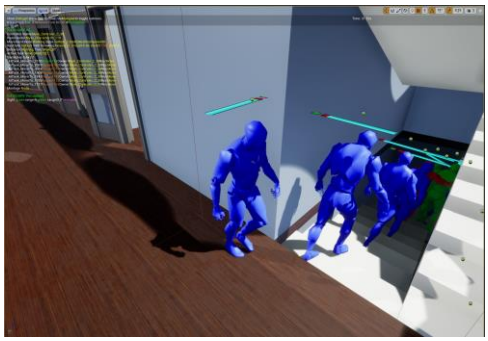
The transitions from *Group* state can be described as

$$T_{4\to5} = \{\langle S_4, leader, leader = 1, \ , S_5\rangle\};$$
$$T_{4\to3} = \{\langle S_4, (leader, stress), leader = 0 \,\&\, \max > Stress > \min, \ , S_3\rangle\}.$$

4 AN EXAMPL OF GRAPHS IMPLEMENTATION

This section provides examples of graphs implementation for changing the physical and psychological states of characters in our model. The practical implementation was carried out in the *Unreal Engine* program.

As noted above when receiving various injuries characters can change their physical state. For example, fall to the *Injury*, *Intoxication* or *Death* states. Figure 13 shows model examples of characters actions in these states.



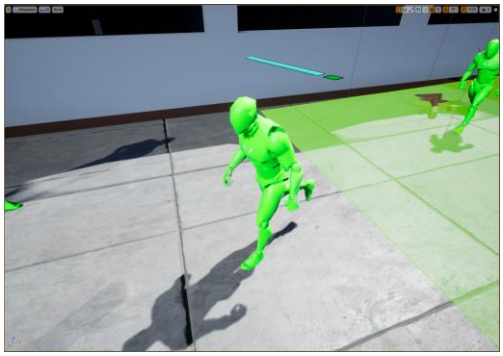
a)

b)

c)

Figure 13: Examples of characters actions in different physical states: a) *Intoxication*; b) *Injury*; c) *Death*.

In model the characters color changes depending on their state. According to the graph shown in Figure 7 the characters can be in various psychological states. These states will determine the actions that the character performs during the evacuation. Figure 14 shows examples of *Calm* and *Sympathy* states.



a)

b)

Figure 14: Examples of characters actions in different psychological states: a) *Calm*; b) *Sympathy*.

In Figure 14a the character is in a *Calm* state and performs the run action at maximum speed. At the same time his speed is determined by the level of *health* and the absence of obstacles. Figure 14b shows the process of providing assistance. In this case another character with a low level of *health* falls into the field of view of the character who is in the *Sympathy* state (red color). Providing assistance leads to their joint evacuation, thereby increasing the speed of the character.

5 CONCLUSIONS

This paper presents the results of the development of graphs (extended finite state machines) describing the physical and psychological states of people in the evacuation process. When developing graphs five states of the characters in the evacuation process were identified. Based on these graphs algorithms for transitions between states were developed and software in the *Unreal Engine* system was implemented. The results of the development formed the basis of a comprehensive 3D modeling system for the people evacuation in emergency situations. In this case we considered a fire in the building as a possible emergency situation. The conducted test experiments on modeling show the adequacy of the choice of states and the set of the characters actions.

In our opinion, the developed graphs and software will increase the realism of modeling processes. Taking into account a large number of states and actions of characters in the model will expand its possible use by specialized organizations. However, it should be noted that the issues of the psychological state of people (characters in the 3D modeling system) and their behavior are quite complex and ambiguous. In this regard, it is planned to conduct additional studies of the adequacy of the simulation results including the expert assessments of psychologists and rescuers.

The development of a detailed 3D model with increased realism of the characters' behavior will open up new areas of such models application. In particular, a detailed accounting of the characters behavior depending on their psycho-physical states will allow to use this model in the field of life safety and obtaining at a new level of detailing statistical data on evacuation processes.

ACKNOWLEDGMENTS

This work was supported by a grant from the Innovation Promotion Foundation (project manager – Shulga Irina).

REFERENCES

[1] D. Helbing, I. Farkas, and T. Vicsek, “Simulating dynamical features of escape panic,” Nature, vol. 407, pp. 487-490, September 2000.

[2] Y. Niu, Y. Zhang, J. Zhang, and J. Xiao, “Running Cells with Decision-Making Mechanism: Intelligence Decision P System for Evacuation Simulation,” International Journal of Computers, Communications & Control (IJCCC), vol. 13, pp. 865-880, September 2018.

[3] J. Kou, Sh. Xiong, Zh. Fang, X. Zong, and Zh. Chen, “Multiobjective Optimization of Evacuation Routes in Stadium Using Superposed Potential Field Network Based ACO,” Computational Intelligence and Neuroscience, vol. 2013, pp. 1-11, 2013.

[4] Y. Wu, J. Kang, and C. Wang, "A crowd route choice evacuation model in large indoor building spaces," Frontiers of Architectural Research, vol. 7, pp.135-150, 2018.

[5] P. Du, Y. Li, H. Liu, and X. Zheng, "Study of the indoor evacuation based on the grouping social force model," 9th International Conference on Information Technology in Medicine and Education, pp. 1018-1026, 2018.

[6] A. Templeton, J. Drury, and A. Philippides, "From Mindless Masses to Small Groups: Conceptualizing Collective Behavior in Crowd Modeling," General Psychology, vol. 19, pp. 215-229, 2015.

[7] H. Liu, B. Liu, H. Zhang, L. Li, X. Qin, and G. Zhang, "Crowd evacuation simulation approach based on navigation knowledge and two-layer control mechanism," Information Sciences, vol. 436-437, pp. 247-267, 2018.

[8] J. Shi, A. Ren, and C. Chen, "Agent-Based Evacuation Model of Large Public Buildings Under Fire Conditions," Automation in Construction, vol. 19, pp. 338-347, 2009.

[9] W. Xin-quan and W. Jian, "A mesoscopic evacuation model based on multi-agent and entropy with leading behavior under fire conditions," Systems Engineering - Theory & Practice, vol. 35, pp. 2473-2483, December 2014.

[10] J. Jumadi, A. J. Heppenstall, N. S. Malleson, S. J. Carver, D. J Quincey, and V. R. Manville, "Modelling Individual Evacuation Decisions during Natural Disasters: A Case Study of Volcanic Crisis in Merapi, Indonesia," Geosciences, vol. 8, p. 196, 2018.

[11] M. A. Tugarinov, I. D. Shulga, E. A. Yurchenko, and A. D. Ermakov, "3D-simulation of emergency evacuation," Journal of Physics: Conference Series, vol. 1680, pp. 1-8, 2020.

[12] M. A. Tugarinov, I. D. Shulga, E. A. Yurchenko, and S. N. Torgaev, "Development of elements of a 3D emergency evacuation simulation system," Journal of Physics: Conference Series, vol. 1680, pp. 1-8, 2020.

[13] M. L. Gromov and N. V. Shabaldina, "Derivation of the cascade parallel composition of timed finite state machines using BALM-II," Automatic control and computer sciences, vol. 51, no. 7, pp. 507-515, 2017.

[14] M. L. Gromov, S. A. Prokopenko, N. V. Shabaldina, and A. V. Laputenko, "Model Based JUnit Testing," 20th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, pp. 139-142, 2019.