Box-Trainer Assessment System with Real-Time Multi-Class Detection and Tracking of Laparoscopic Instruments, using CNN

Fatemeh Rashidi Fathabadi, Janos L. Grantner, and Ikhlas Abdel-Qader

Department of Electrical and Computer Engineering, Western Michigan University, 1903 West Michigan Ave, Kalamazoo, MI 49008, USA
Fatemeh.rashidifathabadi@wmich.edu; janos.grantner@wmich.edu; ikhlas.abdelqader@wmich.edu

Saad A. Shebrain M.D.

Department of Surgery, Homer Stryker M.D. School of Medicine, Western Michigan University, 300 Portage Street, Kalamazoo, MI 49007, USA
saad.shebrain@med.wmich.edu

Abstract: In Minimally Invasive Surgery (MIS), surgeons need to acquire a specific set of skills, before carrying out a "real" operation. Training with the Laparoscopic Surgical Box-Trainer device helps in acquiring the needed skills for surgery residents which are traditionally not taught to them. Video recording of residents' performance and computer-assisted surgical trainers for MIS provide valuable information for resident's assessment. In this paper, we propose real-time detection and tracking of a multi-class of laparoscopic instruments for an intelligent box-trainer performance assessment system using SSD-ResNet50 V1 FPN architecture in TensorFlow backend. The dataset has been extracted from various laparoscopic box training videos. Using distance measurements and evaluation criteria constraints, we present an evaluation of the surgeon's performance. Based on the experimental result, the trained model could identify each instrument at the score of 90% fidelity, in each location, within a region of interest. This research is a result of a partnership between the Department of Electrical and Computer Engineering and the Department of Surgery, of the Homer Stryker M.D. School of Medicine, at Western Michigan University.

Keywords: Intelligent Laparoscopic Surgical Box-Trainer; Laparoscopic Surgical Tool Tip Tracking; Fuzzy Logic-Based Performance Assessment System

1 Introduction

Minimally invasive surgery (MIS) reduces complications and health risks with respect to traditional (open) surgery and decreases hospital stay. However, surgeons must acquire many skills before carrying out a real operation, for example, development of excellent eye-hand coordination while operating using visual information from two-dimensional monitor images and having confident control of the graspers and other laparoscopic surgery instruments, just to mention a few [1] [2]. To help in acquiring these skills various Virtual Reality (VR) trainers have been developed which can assist surgery interns to improve their skills [3] [4]. The assessment of surgical skills requires a considerable amount of time and effort. In recent decades, various training methods have been introduced to provide valuable feedback, expedite the development of surgery skills and assess the trainees' performance [5] [6] [7]. By monitoring a recorded video of a surgeon's performance or observing it in real-time in an Operating Room (OR) during laparoscopic procedures, the assessment procedure can be implemented. Furthermore, object detection and distance estimation concerning the laparoscopic surgical instruments and the test platforms are two fundamental factors for creating an intelligent performance assessment system in MIS [8]. In this paper, we propose a multi-class, real-time detection and tracking system for laparoscopic instruments using SSD-ResNet50 V1 FPN. It will enhance the capabilities of our intelligent box-trainer system [9] [10]. The paper is organized as follows: Section 2 reviews related work in this research area. Section 3 presents tools and utilities which were applied in this approach. Section 4 provides a detailed explanation of the methodology employed in this research. Sections 5 and 6 describe the tracking point location procedure and the model training and evaluation processes, respectively. In Section 7, the real-time tracking and assessment procedures are outlined. Finally, in Section 8, conclusions and plans for further research are given.

2 Related Works

In this section, we focus on the most recent work of researchers regarding surgeons' performance accuracy enhancement during MIS. As mentioned previously, laparoscopic instrument detection and tool-tip tracking contribute to the surgeon's performance assessment. In what follows, we review modern methods that have been proposed in the areas of object detection algorithms, tooltip tracking, and performance assessment [6-35]. Although in the field of MIS, researchers have proposed to apply texture features, color detection, Haar wavelets, and gradient-based features for both processing medical images and hardware-based simulators, there is an emerging trend in recent decades to utilize deep learning approaches [11]. For laparoscopic box-trainer systems, the approach to detect the surgical tools and the movements of tooltips in 3D space by using deep learning along with real-time performance assessment is relatively new.

In more recent studies scholars have been predominantly working on Deep Learning algorithms and Computer-Aided Diagnosis (CAD) system [12]. Yamazaki et al. [6] applied the open-source neural network platform YOLOv3 to detect the movements of surgical instruments in video recordings of laparoscopic gastrectomy procedures.

Namazi et al. [8] proposed a method to assess the surgeon's performance using a Deep Learning System (SPD-DLS) to identify the surgical phases from recorded videos of a laparoscopic procedure. They used a deep Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) model to consider both spatial and temporal information to identify the surgical phases in the video. Grantner et al. in [9] proposed an Intelligent Box-Trainer System (IBTS) to implement tooltip tracking tasks in 3D space, measurement of the forces applied by the grasper's jaws, and task execution time, and an assessment system for the laparoscopic surgeon's performance using fuzzy logic. They worked with a color-filtering algorithm for tool-tip tracking. Allen et al. [13] estimated the tooltip position by detecting the laparoscopic instrument's shaft in each image. They employed color space analysis to extract the instrument contours and then utilized line fitting to estimate the direction of movement for each laparoscopic instrument. In the end, to identify the position of the tool-tip of each instrument they employed a linear search. Perez-Escaminosa et al. [14] detected and tracked movements of laparoscopic instruments in a three-dimensional workspace using a sensor-free system based on green and blue color markers which were placed on the tip of the instruments [13] [15].

In [16], researchers developed a tracking algorithm using a sequence of image contrast enhancement, Sobel Filtering, and color-based segmentation. This algorithm extracts information obtained from the laparoscopic instrument's shaft edge to extract the motion fields of laparoscopic instruments via video tracking. Sun et al. [17] utilized an adaptive fusion kinematics method in an autonomous surgical instrument detection and tracking algorithm. They developed a fuzzy logic system to adjust the kinematics weights and laparoscopic information. Huang et. al. [18] proposed a method to estimate the position, velocity, and direction of laparoscopic instruments which were used in a tracking module. They utilized an Inertial Measurement Unit (IMU) providing direct motion information for the laparoscopic instrument tracking module. Moreover, an Extended Kalman Filter was employed to integrate the information from the different sources to compensate for the biases of the IMU in a unified framework.

Zahiri et al. [19] implemented an Image-Based Tool Tracking system using two-color markers placed on two graspers. Gautier et al. in [20] proposed a surgeon's performance assessment system using colored tapes attached to the end of the laparoscopic instrument. By tracking the colored tapes, frequency analysis and linear discriminant analysis of the 3D reconstructed trajectories of the instruments were extracted to assess the surgeon's skills. Partrige et al. [10] utilized a color-thresholding motion-tracking program to track the movement of colored laparoscopic instrument tips, providing objective performance feedback to a

portable laparoscopic box simulator. Dockter et al. [21] developed a 3D tracking algorithm at a high rate of computational speed to validate its performance in a da Vinci surgical endoscope. Islam et al. [22] presented a tool-tip tracking algorithm based upon a fuzzy logic assessment system by utilizing a web-based video telemonitoring system to monitor and track movements of surgeons' hands and the surgical tooltips. To assess the surgeon's skills, they used colored tool-markers to extract velocity, acceleration, jerkiness, and snaps of the tools' movements during laparoscopic procedures. In recent studies of laparoscopic surgical operations, researchers have been interested in working with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [23]. Using Region-Based Convolutional Neural Networks and a new dataset m2cai16-tool location, Jin et al. [24] used learning of instrument regions in cholecystectomy to detect and localize the region of interest of surgical tools in laparoscopic surgical videos.

In [25], Kletz et al. worked with a Deep Learning instance segmentation approach in recorded videos using a region-based Fully Convolutional Network. They managed to identify instruments as multi-class instance segmentation and determined each instrument classification. Zhang et al. [26] proposed a Modulated Anchoring Network for the detection of laparoscopic surgery tools based on the Faster R-CNN which was made up of a new anchoring scheme referred to as modulated anchoring and a relation module on an existing dataset (m2cai16-tool-locations) as well as a new private one (AJU-Set). Choi et al. [27] proposed real-time models for the detection of surgical instruments during laparoscopic surgery using a dataset that included information on the seven surgical tools for learning the CNN model. To track surgical instruments in real-time, the unified architecture of YOLO was applied to the models.

Wang et al. [28] proposed a multi-label classification deep learning method that combined two deep neural networks, VGGNet and GoogLeNet, to detect the surgical tools in laparoscopic videos. Colleoni et al. [29] proposed a Fully Convolutional Neural Networks (FCNNs) encoder-decoder architecture for surgical instrument joint detection and localization using three-dimensional convolutional layers to exploit spatio-temporal features from laparoscopic videos. The researchers used the EndoVis and UCL dVRK datasets for training testing procedures. Hasan et al. [30] presented a U-NetPlus model for the surgical tool segmentation which is the modification of the U-Net architecture by introducing both VGG-11 and VGG-16 as an encoder and redesigned the decoder part by replacing the transposed convolution operation with an up-sampling operation based on the nearest-neighbor (NN) interpolation followed by two convolution layers.

In the paper by Kurian et al. [31], researchers used the CNN architecture ResNet50 to recognize four surgical phases:

1) Preparation
2) Trocar placement
3) Clipping and cutting
4) Gallbladder retraction

They combined ResNet and temporal features in the form of I3D and LSTM. Kanakatte et al. [32] presented a deep architecture, in which a pixel-wise instance segmentation algorithm segmented and localized the surgical tool in cholecystectomy surgery videos. Jo et al. [33] presented a detection and classification surgical instruments algorithm in laparoscopic images which can work under real-time conditions, too. This algorithm is based on the object detection system YOLO9000. In the paper by Jonmohamadi et al. [34], the researcher used trained fully convolutional neural networks with U-net and U-net++ architectures to segment four key structures of the knee, such as Femur, ACL, Tibia, and Meniscus, in an automated fashion. Zhang et al. [35] developed a marker-free surgical instrument tracking framework based on object extraction using the LinkNet-18 network architecture which belongs to U-Net. In this work, the researchers used a masking method to segment each part of a laparoscopic instrument such as the end-effector, the shaft, and also the background. For real-time tracking, a target trajectory has been defined for the laparoscope-holder robot to be tracked. Using Euclidean Distance Transformation, the binary image was transformed to a distance.

Zijian et al [36] proposed an algorithm that tracks two parts of the surgical instrument: the end-effector and the shaft. In this approach, the shaft detection has been done by edge-points and line features and the trained CNN has been utilized to track and detect the end-effector. Zhu et al [37] proposed an end-to-end learning-based approach to predict distances for given objects in the RGB images. Their method includes three components: a feature extractor, a distance regressor and a multiclass classifier. In this method, a base model extracts features from images, then uses ROI pooling to generate a fixed-size feature vector for each object, and finally feeds the ROI features into a distance regressor to predict the distance for each object.

3 Tools and Utilities

In our study, we propose an intelligent box-trainer performance assessment system based upon real-time detection and tracking of multi-class of laparoscopic instruments. To detect and track the laparoscopic instruments, we used a deep learning approach. Our network is based on an open-source Tensorflow Object Detection Application Programming Interface (API[1]), and we used SSD-ResNet-50 [38] model Feature Pyramid Network (FPN) Architecture as a backbone of our network. TensorFlow is a frequently used software for Machine Learning (ML) applications that provides an interface to common ML algorithms and executable code for various models [39]. In this work, Tensorflow is the backend for object detection and image processing algorithm.

[1] Application Programming Interface (API)

3.1 Experimental Setup

Our Intelligent Box-trainer System (IBTS) is depicted in Figure 1. It is our development platform to create hardware and software architectures and algorithms aiming for the development of an objective laparoscopic-surgery skills assessment system [9]. The main components of the IBTS are as follows: an FLS box-trainer device, two 5-megapixel USB 2.0 cameras with variable-focus lens, a 32' HD computer monitor to visualize the underlying test procedure carried out on a particular FLS platform, extra LED strips for better lighting conditions, a tablet which is used by the supervising medical personnel, a PC workstation to record the test videos and run the tracking and assessment programs and a router to implement wireless communications between the tablet and the PC workstation. One of the standard FLS pattern cutting tests was used in this study. In this test, the platform is an artificial tissue with a bold circle line on it. During the test, the surgeon holds the tissue in place by a grasper while using a pair of scissors to cut the tissue around the circle such that the cutting distance from the circle line should not exceed 5 mm, and cutting into the line is also considered as a failure of the test.

(a) (b)

Figure 1

a) The IBTS System

b) Tracking the laparoscopic instruments and generating real-time performance assessment using the IBTS System in the Homer Stryker M.D. School of Medicine, of WMU

3.2 Dataset

In this study, we used our custom dataset (IFCL.LBT100) that has been created for laparoscopic box trainer's performance assessment research. For this project, we have created a relatively large dataset using various laparoscopic training videos. Our custom dataset is composed of extracted frames from these videos. The frames have been manually annotated using the Image Annotation Tool LabelImg[2] which

[2] http://tzutalin.github.io/labelImg/

is a free and open-source software. Each labeled image has its individual .xml which can be converted to .csv files and .tf.record files which are used during training processes. Having recorded and processed more test videos we plan to post our dataset online for other researchers in this field.

3.2.1 Partitioning of the Dataset

Once the annotation of our dataset was finished, we classified all the images and .xml, .csv, and .tf record files for the training, testing, and evaluation tasks. Typically, the ratio of this arrangement is 6:2:2, i.e., 60% of the images are used for training, 20% for testing, and the remaining 20% is used for evaluation purposes.

3.2.2 Creating the Label Map

To satisfy the training algorithm, we prepared a label map that maps each of the classes to an integer value. This label map is used both by the training and detection processes. A simple example of the label map for our dataset contains three labels: a scissor, a grasper, and a circle pattern on an artificial tissue which are considered as the laparoscopic instruments in a laparoscopic box-trainer. This label map file (with the extension of .pbtxt) is illustrated in Figure 2.

```
item {
    id: 1
    name: 'Scissor'
}

item {
    id: 2
    name: 'Grasper'
}
item {
    id: 3
    name: 'Circle'
}
```

Figure 2
Label map file example

3.2.3 Data Augmentation

The TensorFlow Object Detection API Image Preprocessor tool provides multiple data augmentation steps with variation and modification from the original data. Applying these augmentation steps to the dataset the neural networks can use more training data to achieve better performance. In our approach, to train the model based on SSD-ResNet50 V1 FPN, we adequately augmented the dataset using TensorFlow API data augmentation variables.

4 Methodology

We selected a collection of detection models and pre-trained them on the COCO 2017 dataset such as the EfficientDet D1 640x640, SSD MobileNet V1 FPN 640x640, and SSD-ResNet50 V1 FPN from TensorFlow 2 Detection Model Zoo and Detecto Module in Pytorch [40]. These models are useful for initialization when training on our new datasets. By comparing the performance of these models, we have concluded that SSD-ResNet50 delivers better performance with respect to real-time detection. We trained our model based upon the SSD-ResNet50 V1 FPN Architecture. The entire workflow of the SSD-ResNet50 V1 FPN Architecture is illustrated in Figure 3. SSD with the ResNet50 V1 FPN feature extractor in its architecture is an object detection model that has been trained on the COCO 2017 dataset. A Momentum optimizer with a learning rate of 0.04 was used for the region proposal and classification network, and the learning rate was reduced on the plateau. As shown in Figure 3, the Feature Pyramid Network (FPN) generates the multi-level features as inputs to the SSD-ResNet50 Architecture. The FPN is an extractor and provides the extracted feature maps layers to an object detector. When the model localizes any small object, it draws an object boundary box around it at each location. After training the model, the testing procedure was carried out by providing the surgical videos as input to the trained model. Afterward, we used Tensorboard which is a suitable feature of the TensorFlow Object Detection API. It allowed us to continuously monitor and visualize several different training/evaluation metrics when our model was being trained. As the final step, we obtained the output video containing the labeled surgical instruments and the assessment results along with the log file. The generated log file records the surgical assessment, the bounding box for each laparoscopic instrument, and the center point of each laparoscopic instrument.

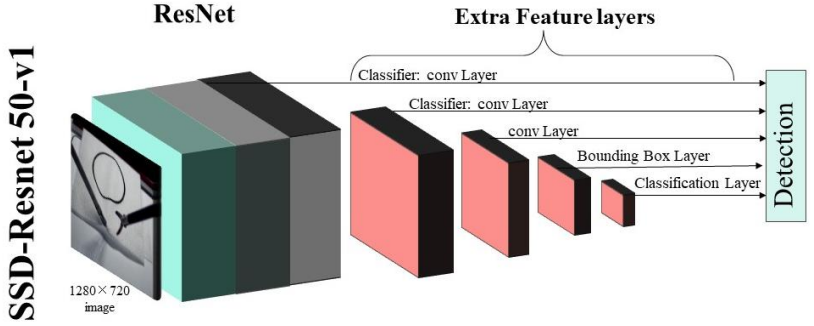


Figure 3
SSD-ResNet50 V1 FPN Architecture

5 Tracking Point Location Procedure

By providing the surgical videos as the input to the trained model, we detected and localized each laparoscopic instrument with high accuracy. After object extraction, we used Euclidean Distance Transformation[3] [35] to measure the distance between the center of the circle pattern and the center of the scissors' bounding box where the tissue is cut. To assess the surgeon's performance, we measure the distance between the spot where the tissue is cut and the closest area of the circle. This distance should not exceed 5 mm for passing the assessment test. The measurement procedures and formulations are defined by Eqs. (1) thru (8). Dis stands for the distance of two points in each frame and Pix stands for the pixel point set which contained all the pixel points of an extracted object of each frame.

$$D[A][B] = \min\{Dis[(A_x . A_y) . (B_x . B_y)] . \ (A.B) \in Pix\}$$
$$Dis[(A_x . A_y) . (B_x . B_y)] = \sqrt{(B_x - A_x)^2 + (B_y - A_y)^2} \quad (1)$$

Given a line $y = mx+b$, the slope m delineates the ratio between the change in x, defined by dx, and the change in y, defined by dy. Hence, the slope creates a relationship between a change in the y-values with respect to a change in the x-values which is a derivative of y to x:

$$dy = mdx \quad (2)$$

$$dy^2 = (B_y - A_y)^2 \quad (3)$$

$$dx^2 = (B_x - A_x)^2 \quad (4)$$

$$dx^2 + dy^2 = (B_x - A_x)^2 + (B_y - A_y)^2$$
$$dx^2 + dy^2 = Dis^2 \quad (5)$$

By substituting Eq. (2) in Eq. (5) we obtain Eq. (6)

$$dx^2 + (mdx)^2 = Dis^2 \quad (6)$$

Using Eq. (2) and Eq. (6), we can calculate the changes in the x-values by Eq. (7) and the y-values by Eq. (8) to obtain the distance between each of two points in our approach.

$$dx = \sqrt{\frac{Dis^2}{1 + m^2}} \quad (7)$$

[3] https://github.com/alejandrods/Social-Distance-Using-TensorFlow-API-Object

$$dy = m\sqrt{\frac{Dis^2}{1+m^2}} \tag{8}$$

Using these equations, we measured the bounding boxes of each laparoscopic instrument in each frame. Based upon real measurements and the number of pixels in each frame, we calibrated the position of each instrument to a real value. The assessment measurement algorithm is illustrated in detail in Figure 4. There, A marks the center of the circle pattern, and B marks the center of the scissors box. Line d, which connects points A and B is defined as the distance of the scissors from the center of the circle. Using Eq. (1) to Eq. (8), we calculated the radius of the circle in each frame.

By subtracting the center of the scissors bounding box from the center of the circle pattern, the distance between the scissors and the center of the circle is calculated for the assessment procedure. These calculations may lead to inaccurate measurements under some circumstances, e.g., when the grasper wrinkles the artificial tissue, or when the trained model cannot recognize the instrument. In our research, there were some short periods, typically lasting for a few seconds, when the model couldn't find the circle which, lead to inaccurate measurement. In other cases, the model could detect the circle by keeping the reference center of the circle in its place by localizing the bounding box of the circle in each frame.

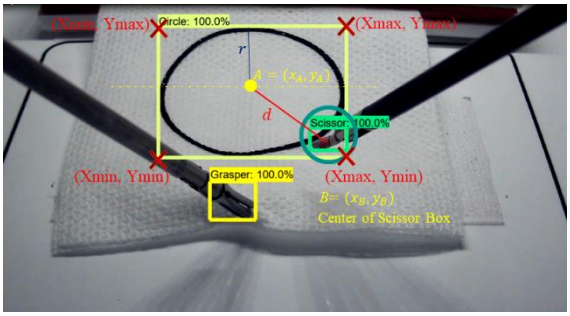


Figure 4
Illustration of the assessment measurement calculations

6 Model Training and the Evaluation Process

The classification loss, which is used to measure the model's confidence by classifying the pixel's region confined by the bounding box, is illustrated Figure 5. The localization loss that measures the geometric distance between the predicted bounding box and the ground truth annotation (validation bounding boxes) is depicted in Figure 6. The overall loss function or total loss is a weighted

combination of the classification loss and the localization loss [41]. It is depicted in Figure 7, which illustrates the performance of the model during training, i.e., what the network predicts for the image versus the allocated label of the image at the end of each epoch during the training process. The train-validation total loss, as it is shown Figure 7, is sometimes higher than the training loss but it decreases over time and, hence, it exhibits a satisfactory result.

Figure 5
The comparison of overall train-classification loss and train-validation classification loss

Figure 6
The comparison of overall train- localization loss and train-validation localization loss

Figure 7
The comparison of overall train-total loss and train-validation total loss

7 Real-time Tracking and the Assessment System

To implement the tracking task, the tracking point has to be located frame-by-frame in the laparoscopic test videos. In our implementation, we analyzed tracking and generated the surgeons' performance assessment along with it.

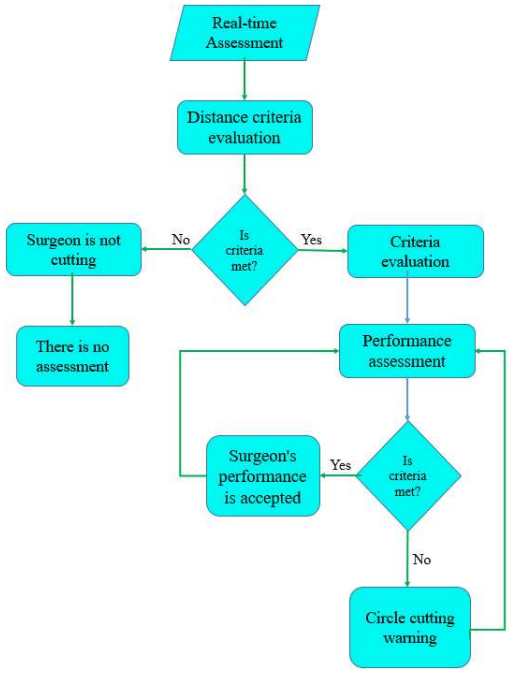
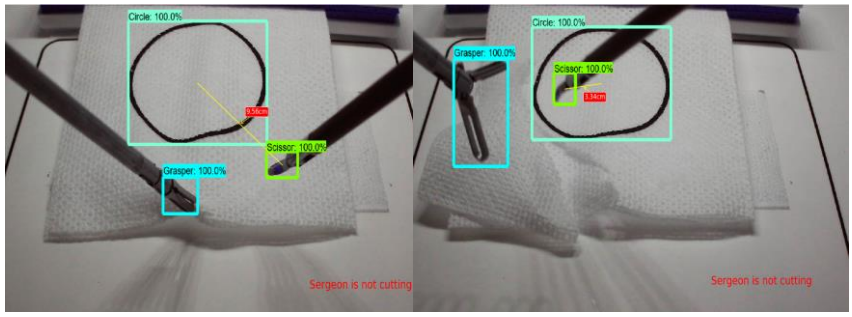


Figure 8
The real-time tracking and assessment system flowchart

In the tracking procedure, we expect that the network detects and localizes each laparoscopic instrument in real-time. However, the surgeon's performance will be assessed only during the actual cutting of the tissue. To clarify the reasons for this, we have simplified the assessment procedure to just three parts. In each part, many components are taken into account for the surgeon's performance assessment. Only the following processes are considered: (1) the surgeon is not cutting, (2) criteria evaluation and performance assessment, and (3) circle cutting warning if the established criteria are not met. The tracking and assessment system flowchart is depicted in Figure 8. In what follows, we investigate each of these processes.

7.1 Surgeon is not Cutting and the Procedure does not Commence

In this situation, either the surgeon has not started the cutting or the scissors are in the air, i.e., they are way above the artificial tissue. By defining different constraints for the distance of the center of the circle to the tips of the scissors, the network can recognize when the surgeon is not cutting. In Figure 9 (a to h), different conditions are illustrated when no cutting takes place. When the surgeon is about to start the cutting (a), there is no need for performance assessment. When the surgeon is in the middle of the cutting process but he temporarily stops doing it and releases the tissue to take a different approach for continuing the task, there is no cutting, either. The network recognizes it when a surgeon is not cutting (b, d, e, h). In two illustrated scenarios the surgeon intends to continue with the cutting but just a small section of the scissors is in the frame. In these cases, the network correctly recognizes the situation and decides "no cutting" is taking place (f, g).



(a) (b)

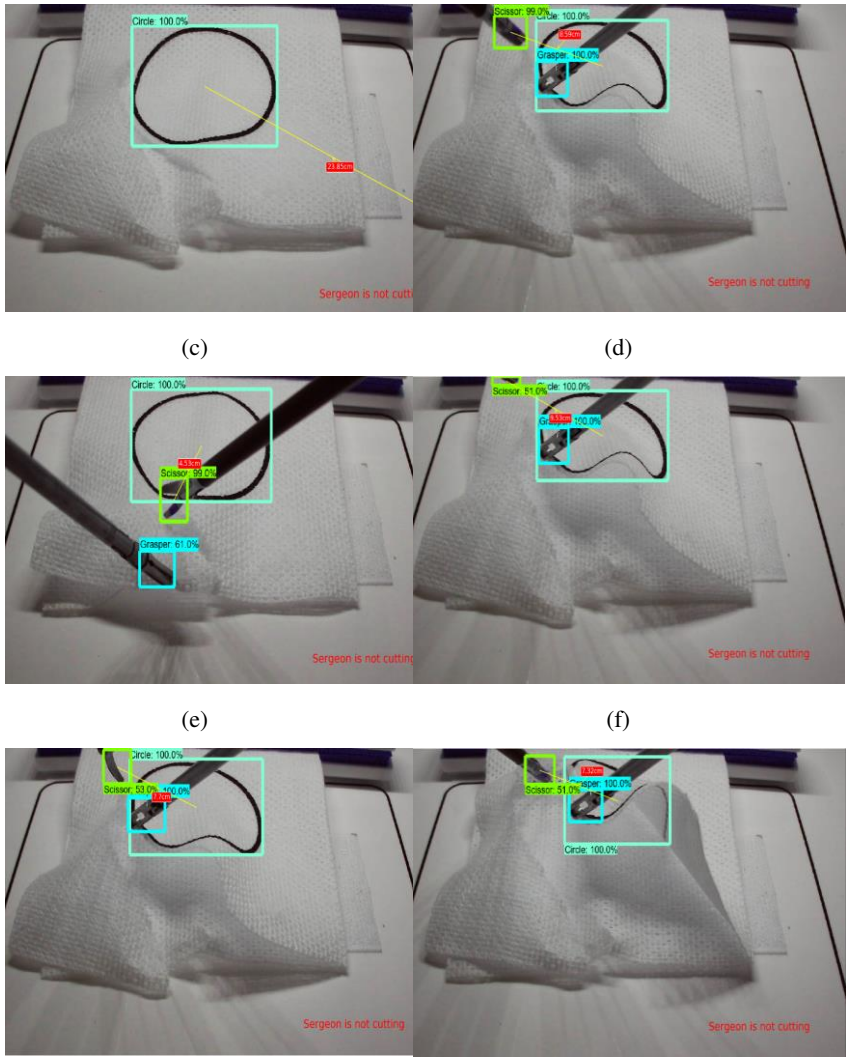


Figure 9
Surgeon is not cutting leading to no assessment

7.2 Performance Assessment is Active

The surgeon's performance will be assessed based upon an error distance, d_e, a distance between the spot where the scissors made the cut and the section of the circle which is at the closest point to the scissors' tips, as shown in Figure 10. The surgeon's performance is not acceptable unless the error distance d_e<5 mm, at

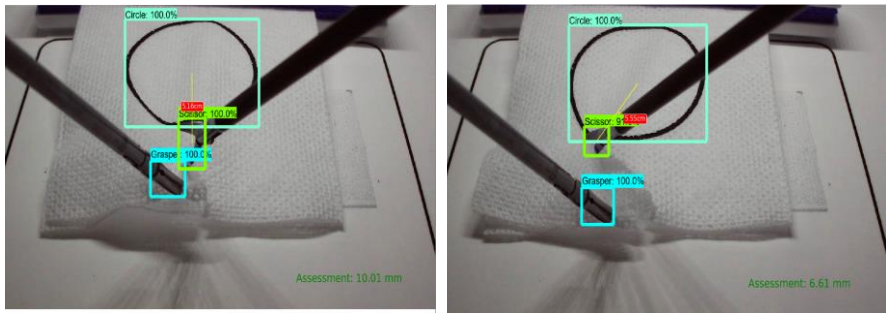
all times. If this condition is not met, the system renders an assessment but it doesn't accept the surgeon's performance. In this case, there are three scenarios to consider.

1) In Figure 10 (a), the surgeon is cutting, however, he is only getting to a position in which he can actually begin with the task. The tips of the scissors are too far from the circle and the network should only deliver assessment during the pattern cutting task. Clearly, in this situation, we do not expect the network to deliver any valid assessment.

2) In Figure 10 (b), the tips of the scissors are close enough to the circle, so after this moment, the surgeon's performance will be monitored and assessed. Based upon Figs. 10 (c, d, e), the surgeon's performance is deemed good because the measured distance between the tips of the scissors and the circle line is less than 5 mm. In Figure 10 (f), the surgeon restarted cutting after he had changed the direction he wanted to move the scissors. As expected, in the case of each start, the distance of the scissors' tips from the circle line is typically larger than that when the surgeon is cutting continuously.

3) In Figs. 10 (g, h), the most challenging scenarios are illustrated: the tissue is wrinkled by the grasper but the reference center of the circle is still in its place, i.e., it is visible to the camera.

Having enough images in a dataset has a great impact on training a model. In our study, because the number of images was not as large as it should be, the model could not recognize the circle in few instances. In particular, when the tissue is wrinkled by the grasper such that the circle line disappears from the sight of the camera. In a situation like that not only the complete circle cannot be recognized by the model, but even an expert cannot recognize it as a circle. Therefore, in such moments, we have a pass-fail assessment. To address this problem, we have to record more videos including many frames of such cases intentionally containing this scenario. Under normal conditions, this situation rarely happens. The more we can train a model to understand this scenario, the better the prediction analysis will be. In addition, installing a third camera into the system, which is positioned directly above the platform, will help in resolving this problem, as we continue our research.



(a) (b)

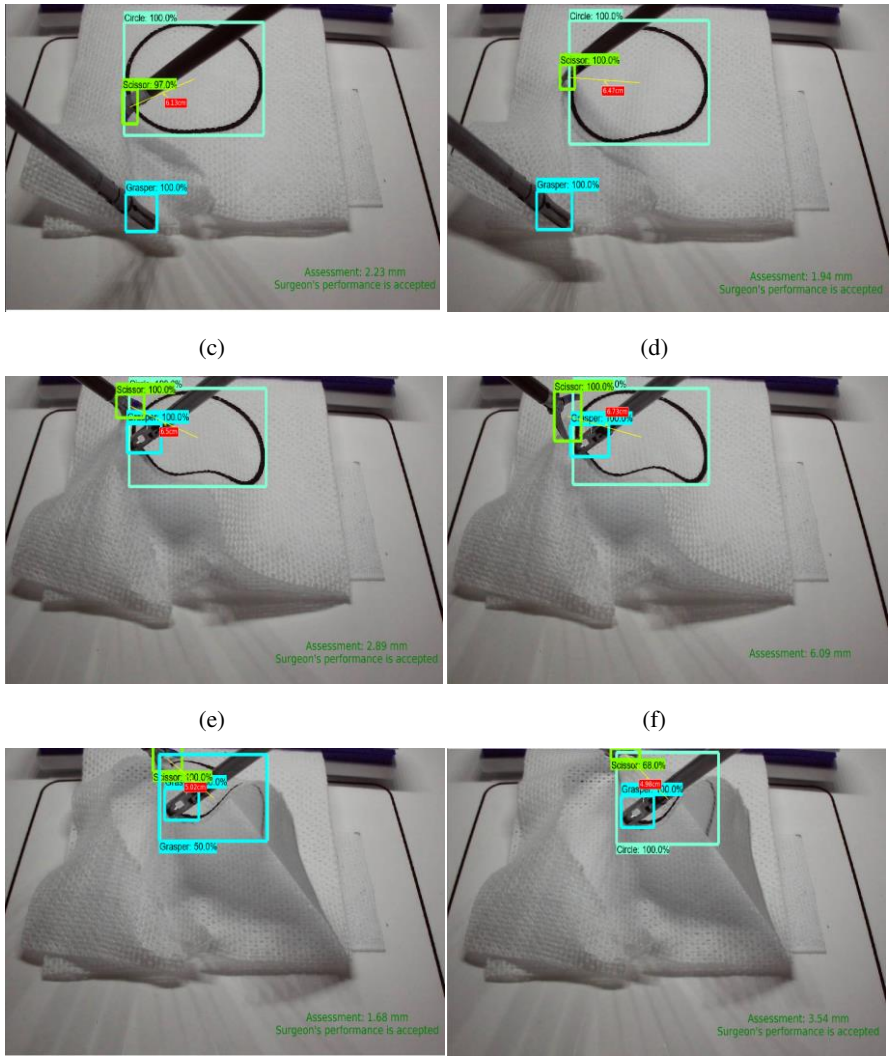


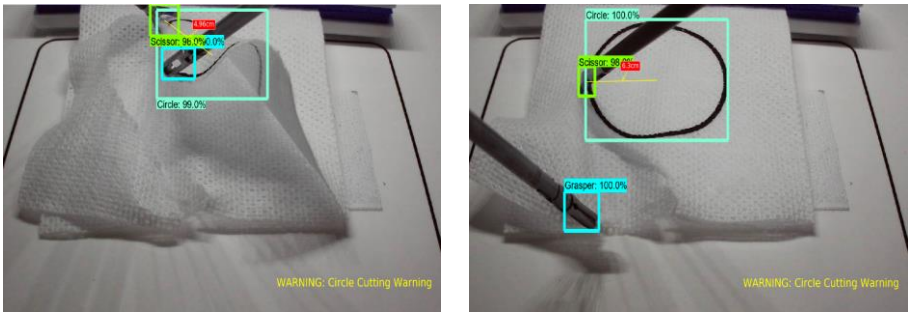
Figure 10

Surgeon's performance is assessed leading to performance acceptance or rejection

7.3 Circle Cutting Warning and Criteria is not Met

There is an additional requirement for passing the pattern cutting test. It is mandatory that the surgeon should not cut through the circle line. To help in meeting this constraint, we defined a rule in which if the tips of the scissors are too close to the circle line (defined as less than 0.5 mm) it will alert the surgeon about this

situation. This scenario is illustrated in Figure 11, when the surgeon was cutting too close to the circle line.



(a) (b)

Figure 11

Surgeon is getting too close to the circle, leading to circle cutting

Conclusions and Future Research

In this paper, we proposed employing real-time detection and tracking of a multi-class of laparoscopic instruments for an intelligent box-trainer performance assessment system. We generated the dataset using extracted frames of various training videos using a laparoscopic box-trainer. Moreover, we added a distance measurement algorithm to the object detection algorithm in the TensorFlow backend using ResNet-50- architecture. The algorithm continuously measures changes in the distance of the center of the circle from the center of the scissors' tips and also the changes of the distance, where the tissue is been cut, from the circle line. Using distance measurements and evaluation criteria constraints, we assessed whether the surgeon's performance was accepted or not. Based on the experimental result, the trained model could identify each instrument at the score of 90% fidelity, in each location within a region of interest, and determine their relative distance with 65% reliability, under real-time conditions. There were few instances when the detection failed to lead to pass-failed assessment, in particular, when the tissue was wrinkled by the grasper. The error rate in carrying out these tasks was less than 20%. We assume that the performance measures of the system can be improved by adding an additional, top camera to the system and measure the distance from different perspectives. In future research, we plan to develop an automated performance assessment system, by tracking the laparoscopic instruments, under real-time conditions, measuring the test execution times and fusing the measured data with expert surgeon opinion, in the framework of a fuzzy logic-based intelligent decision support system.

Acknowledgement

This work was supported by the Homer Stryker M.D. School of Medicine, WMU (Contract #: 29-7023660), and the Office of Vice President for Research (OVPR), WMU (Project #: 161, 2018-19). We gratefully thank Mr. Hossein Rahmatpour

(University of Tehran) for sharing his experience in this type of work, *hosseinrahmatpour@ut.ac.ir*.

References

[1] A. Chellali *et al.*, "Achieving interface and environment fidelity in the Virtual Basic Laparoscopic Surgical Trainer," *Int. J. Hum. Comput. Stud.*, Vol. 96, pp. 22-37, 2016

[2] D. Oh *et al.*, "Surgical techniques for totally laparoscopic caudate lobectomy," *J. Laparoendosc. Adv. Surg. Tech.*, Vol. 26, No. 9, pp. 689-692, 2016

[3] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surg. Endosc.*, Vol. 32, No. 4, pp. 1636-1655, 2018

[4] H. Jiang *et al.*, "Enhancing a laparoscopy training system with augmented reality visualization," in *2019 Spring Simulation Conference (SpringSim)*, 2019, pp. 1-12

[5] B. W. King, L. A. Reisner, A. K. Pandya, A. M. Composto, R. D. Ellis, and M. D. Klein, "Towards an autonomous robot for camera control during laparoscopic surgery," *J. Laparoendosc. Adv. Surg. Tech.*, Vol. 23, No. 12, pp. 1027-1030, 2013

[6] Y. Yamazaki *et al.*, "Automated Surgical Instrument Detection from Laparoscopic Gastrectomy Video Images Using an Open Source Convolutional Neural Network Platform," *J. Am. Coll. Surg.*, Vol. 230.5, pp. 725-732, 2020

[7] S. Ahmadi and S. J. Azhari, "A LP, Very High-CMRR, Wide-Bandwidth FDCCII-Based CMIA Adapted to Both Current and Voltage Inputs," *Arab. J. Sci. Eng.*, Vol. 44, No. 8, pp. 6727-6740, 2019

[8] B. Namazi, G. Sankaranarayanan, and V. Devarajan, "Automatic detection of surgical phases in laparoscopic videos," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2018, pp. 124-130

[9] J. L. Grantner, A. H. Kurdi, M. Al-Gailani, I. Abdel-Qader, R. G. Sawyer, and S. Shebrain, "Multi-Thread Implementation of Tool Tip Tracking for Laparoscopic Surgical Box-Trainer Intelligent Performance Assessment System," *Acta Polytech. Hungarica*, Vol. 16, No. 9, 2019

[10] R. W. Partridge, M. A. Hughes, P. M. Brennan, and I. A. M. Hennessey, "Accessible laparoscopic instrument tracking ('InsTrac'): construct validity in a take-home box simulator," *J. Laparoendosc. Adv. Surg. Tech.*, Vol. 24, No. 8, pp. 578-583, 2014

[11] X.-Y. Zhou, Y. Guo, M. Shen, and G.-Z. Yang, "Artificial Intelligence in Surgery," *arXiv Prepr. arXiv2001.00627*, 2019

[12] P. Eskandari and S. B. Shokouhi, "DT-CWT: A New Feature for Tumor Classification in Breast DCE-MRI," *Mapta J. Electr. Comput. Eng.*, Vol. 3, No. 1, pp. 35-39, 2021

[13] B. F. Allen, F. Kasper, G. Nataneli, E. P. Dutson, and P. Faloutsos, "Visual tracking of laparoscopic instruments in standard training environments.," in *MMVR*, 2011, pp. 11-17

[14] F. Pérez-Escamirosa, I. Oropesa, P. Sánchez-González, J. Tapia-Jurado, J. Ruiz-Lizarraga, and A. Minor-Martínez, "Orthogonal cameras system for tracking of laparoscopic instruments in training environments," *Cir. Cir.*, Vol. 86, No. 6, pp. 548-555, 2019

[15] C. B. Duane, "Close-range camera calibration," *Photogramm. Eng*, Vol. 37, No. 8, pp. 855-866, 1971

[16] I. Oropesa *et al.*, "EVA: laparoscopic instrument tracking based on endoscopic video analysis for psychomotor skills assessment," *Surg. Endosc.*, Vol. 27, No. 3, pp. 1029-1039, 2013

[17] Y. Sun, B. Pan, S. Zou, and Y. Fu, "Adaptive Fusion-Based Autonomous Laparoscope Control for Semi-Autonomous Surgery," *J. Med. Syst.*, Vol. 44, No. 1, p. 4, 2020

[18] C.-C. Huang, N. M. Hung, and A. Kumar, "Hybrid method for 3D instrument reconstruction and tracking in laparoscopy surgery," in *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2013, pp. 36-41

[19] M. Zahiri, R. Booton, K.-C. Siu, and C. A. Nelson, "Design and evaluation of a portable laparoscopic training system using virtual reality," *J. Med. Device.*, Vol. 11, No. 1, 2017

[20] B. Gautier, H. Tugal, B. Tang, G. Nabi, and M. S. Erden, "Laparoscopy instrument tracking for single view camera and skill assessment," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5039-5045

[21] R. Dockter, R. Sweet, and T. Kowalewski, "A fast, low-cost, computer vision approach for tracking surgical tools," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 1984-1989

[22] G. Islam, K. Kahol, B. Li, M. Smith, and V. L. Patel, "Affordable, web-based surgical skill training and evaluation tool," *J. Biomed. Inform.*, Vol. 59, pp. 102-114, 2016

[23] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imaging*, Vol. 36, No. 1, pp. 86-97, 2016

[24] A. Jin *et al.*, "Tool detection and operative skill assessment in surgical videos

using region-based convolutional neural networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 691-699

[25] S. Kletz, K. Schoeffmann, J. Benois-Pineau, and H. Husslein, "Identifying surgical instruments in laparoscopy using deep learning instance segmentation," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1-6

[26] B. Zhang, S. Wang, L. Dong, and P. Chen, "Surgical Tools Detection Based on Modulated Anchoring Network in Laparoscopic Videos," *IEEE Access*, Vol. 8, pp. 23748-23758, 2020

[27] B. Choi, K. Jo, S. Choi, and J. Choi, "Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1756-1759

[28] S. Wang, A. Raju, and J. Huang, "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 620-623

[29] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robot. Autom. Lett.*, Vol. 4, No. 3, pp. 2714-2721, 2019

[30] S. M. K. Hasan and C. A. Linte, "U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 7205-7211

[31] E. Kurian, J. J. Kizhakethottam, and J. Mathew, "Deep learning based Surgical Workflow Recognition from Laparoscopic Videos," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 928-931

[32] A. Kanakatte, A. Ramaswamy, J. Gubbi, A. Ghose, and B. Purushothaman, "Surgical tool segmentation and localization using spatio-temporal deep network," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 1658-1661

[33] K. Jo, Y. Choi, J. Choi, and J. W. Chung, "Robust Real-Time Detection of Laparoscopic Instruments in Robot Surgery Using Convolutional Neural Networks with Motion Vector Prediction," *Appl. Sci.*, Vol. 9, No. 14, p. 2865, 2019

[34] Y. Jonmohamadi *et al.*, "Automatic segmentation of multiple structures in knee arthroscopy using deep learning," *IEEE Access*, Vol. 8, pp. 51853-51861, 2020

[35] J. Zhang and X. Gao, "Object extraction via deep learning-based marker-free tracking framework of surgical instruments for laparoscope-holder robots," *Int. J. Comput. Assist. Radiol. Surg.*, Vol. 15, No. 8, pp. 1335-1345, 2020

[36] Z. Zhao, S. Voros, Y. Weng, F. Chang, and R. Li, "Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method," *Comput. Assist. Surg.*, Vol. 22, No. sup1, pp. 26-35, 2017

[37] J. Zhu and Y. Fang, "Learning object-specific distance from a monocular image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3839-3848

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778

[39] O. Alsing, "Mobile object detection using tensorflow lite and transfer learning" 2018

[40] F. R. Fathabadi, J. L. Grantner, S. A. Shebrain, and I. Abdel-Qader, "Multi-Class Detection of Laparoscopic Instruments for the Intelligent Box-Trainer System Using Faster R-CNN Architecture," in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI 2021)*, pp. 149-154

[41] Y. Wu *et al.*, "Rethinking Classification and Localization for Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10186-10195

Efficiency Improvement of the GLOBAL Optimization Method by Local Search Changes

Abigél Mester[1], Dániel Zombori[1], László Pál[2], Balázs Bánhelyi[1*]

[1] University of Szeged, Institute of Informatics
Dugonics tér 13, H-6720 Szeged, Hungary
mester@inf.u-szeged.hu; zomborid@inf.u-szeged.hu; banhelyi@inf.u-szeged.hu

[2] Sapientia Hungarian University of Transylvania, Faculty of Economics, Socio-Human Sciences and Engineering, Piaţa Libertăţii nr. 1, 530104 Miercurea Ciuc, Romania; pallaszlo@uni.sapientia.ro

[*] Corresponding author

Abstract: There are many suitable global optimization approaches to find the minimum value of an objective function. In this paper, the improvement of the GLOBAL Optimization Method is studied, which is based on stochastic clustering. Through its three main components, which are sampling, clustering, and local search the algorithm aims to find the global minimum of the objective function. Local search methods significantly influence the efficiency of the GLOBAL method. The efficiency of our proposal may be improved by dividing the system into modules and by creating new variants of both the local and line search methods. The main contribution of this work is to show the achievements of modularization and the efficiency of the new variants of both local and line search methods.

Keywords: GLOBAL; Optimization; Local search; Line search; Modular software

1 Introduction

There is a wide range of optimization problems ranging from everyday tasks [1, 2] to economic issues and theoretical chemical problems [3]. Various stochastic, deterministic, and hybrid global optimization methods have been used to solve these problems. For example, in these papers, the authors explore numerical solution techniques for economic and chemical problems, using differential evolution (DE) or genetic algorithm (GA). In recent years, our research group has solved the same problems using GLOBAL. It has already proved its relevance in such diverse problems [4, 5, 6, 7]. Moreover, it even mastered difficult mathematical problems from the field of qualitative analysis of dynamical systems [6, 8, 9,10]. In these works GLOBAL was used to find a feature. For example, we find the parameters of regions that exhibit chaotic behavior.

These problems consist of global and local optimization challenges. During the execution of the global search method, we sweep the entire search space, and local searches are only executed in near areas where potential optimum values can be found. Since global search usually cannot precisely find the global optimum, we have to run it until a stopping criteria is reached. Afterwards, we run local searches, which are based on function evaluations that are mostly very expensive. Therefore, the number of local search runs should be considered. At this point, the clustering is responsible for reducing the number of sample points.

As we start running the global search algorithm, it generates sample points in the search space, which are then clustered around the local optimum's basin of attraction. If a point cannot be added to an existing cluster, we have to start a local search from that point in order to decide which cluster it should be classified to.

As we follow the operation of the search method, we can easily recognize the three main components, which are sampling, clustering, and local search. So we separated, our system into three different modules. This way, any of the algorithms can easily be implemented, improved, or even replaced. By running and testing these modules, we have realized that due to the high number of function evaluations the algorithm spends most of the time executing local searches [11]. Therefore, our aim was to reduce the time spent in local searches. We found two possible methods to achieve our goal. One is to rewrite the local search method, the other is to revise the line search method. Both methods have been implemented and tested.

First of all, we restructured the original Unirandi local search algorithm, so the line search method became an independent module. Afterwards, we created more local search processes, and we were able to develop three different line search approaches based on polynomial interpolation techniques as a result.

2 Environment

2.1 Modularized GLOBAL

Former versions of GLOBAL were available in Fortran, C, and Matlab [5, 12, 13]. To make GLOBAL more efficient to work with an easily extendable optimizer, we needed an object oriented implementation. To achieve that, we implemented GLOBAL in Java and separated the system into the three individual modules described earlier [15]. The two major modules are the clustering and the local search one, and it's the close cooperation of these modules which makes the algorithm efficient.

In the original version, clustering was an integral part of GLOBAL, and the line search was also integrated into the local search method. Due to the modularization,

these are significantly easier to call and customize. During the implementation, we had to ensure problem-free communication between the individual modules. We have solved this by using the Builder pattern.

An advantage of the Builder pattern is that building and parameterization can be automated. An interface is provided where parameters are passed through an XML file and are processed by the Builder methods. A proper Optimizer object holding the specified configuration is created, which can be called by the user through its interface.

2.1.1 Builder Pattern

The Builder design pattern ensures the proper module parameterization and helps easy reporting of misconfigurations. For every module ,we need a *Builder* nested class that implements setter methods for the module parameters and the required sub-modules. The builder also implements the *build()* function to produce a correctly parameterized module instance. In this function, we can check for the setting of required parameters, we can set default values, we can check if incompatible settings are present and we can log the configuration or report problems in a principled way.

2.1.2 XML Configuration

In a former version of the Java implementation, a lot of function calls were required to set the parameters one after the other. Not just the code was hard to read this way, but there were no efficient ways of testing larger sets of functions with different parameters. To solve this, we have developed a simple configuration building system that takes an XML configuration file and automatically generates the necessary function calls. This system automatically exposes all modules and their parameters to the XML file, relying only on the module structure. In this way, plenty of different parameterizations can coexist in XML files. Finally, it is much more readable when we are reviewing our code. The development of a graphical configurator also became much easier.

In the XML sample code below, the construction of the system is a lot more comprehensible. GLOBAL has search parameters such as sample size, and it needs clustering and local search modules. The XML tags correspond to parameters in the configuration tree. The root is the GLOBAL optimizer module, the implementing Java class is defined in the *class* attribute. The tags under Global hold the parameters identified by the tag names. The literal parameters can be one of the *double*, *long* and *string* types. They will result in a function call on their parent module, for example, the *NewSampleSize* parameter will become *moduleBuilder.setNewSampleSize(longValue)*. Sub-modules are similarly easy to define, Global's *LocalOptimizer* parameter will be set to a local optimizer object. The sub-modules Java class is specified in the *class* attribute, the sub-module is

built like the root module, in a recursive way. When a module's parameters are set, the Builder will output a configured object.

```
<? xml version ="1.0" ?>
<Global package =" org.uszeged.inf.optimization.algorithm "
        class =" optimizer.global.serialized.SerializedGlobal ">
        < NewSampleSize type =" long "> 50 </ NewSampleSize >
        < SampleReducingFactor type =" double "> 0.04 </SampleReducingFactor >
        < LocalOptimizer class =" optimizer.local.parallel.NUnirandiCLS ">
                < MaxFunctionEvaluations  type =" long "> 100000
                </ MaxFunctionEvaluations >
                < RelativeConvergence type =" double "> 1e-8
                </ RelativeConvergence >
                < LineSearchFunction class =" optimizer.line.parallel.Fminbnd5 ">
                </ LineSearchFunction >
        </ LocalOptimizer >
        <Clusterizer
        class =" clustering.serialized.SerializedGlobalSingleLinkageClusterizer ">
                <Alpha type =" double "> 0.01 </ Alpha >
        </ Clusterizer >
</ Global >
```

The class attribute of the node points to the Global class in the Optimizer package. The first subnode is NewSampleSize, the type is long and has a value of 50, which means that the optimizer will randomize at most 50 starting points for local search in the search space. The second argument is the SampleReducingFactor with type double and a value of 0.04. This means that 4% of the NewSampleSize (the default value is 100%, or 1) will be selected from the generated sample points. In this case, it is the best 2 samples. The LocalOptimizer node contains a class that must be instantiated. The algorithm is specified in the class argument, which must implement the interface specified by Global's *setLocalOptimizer(LocalOptimizer)* function. MaxFunctionEvaluations is a parameter for the local search module, not the whole optimization process. RelativeConvergence is a termination criterion, smaller values represent a longer and more accurate run. The LineSearchFunction submodule has no parameters itself, but it is a parameter of the LocalOptimizer. The Clusterizer is a parameter of the Global class. The Alpha parameter controls the rate at which the clusterizer shrinks its region of influence. The larger the value (in the interval [0;1] inclusive interval), the harder it is to cluster a sample. At 1 clustering is effectively disabled.

2.2 Local Search

Local search is a crucial part of the GLOBAL algorithm, hence the performance depends a lot on the attached local search method. GLOBAL randomly chooses points in the search space, then, by evaluating them we can conclude some information about the location of the global optimum. We attempt to create clusters

around the local optima from the sample points, which fall into the local optima's basin of attraction. To assign the unclustered points to a cluster, we have to start a local search from the points which could not be assigned to clusters that were already found based on the applied clustering criteria. However, local search is an expensive technique because it takes many function evaluations, so we would like to limit its use.

The easy change of the GLOBAL algorithm parts was the main profit of the modular implementation. The Local search algorithm is one module of the GLOBAL Java implementation. All described local search methods contain a line search technique which can be considered as a separate module. Three kinds of local search and three variants of line search methods were implemented in the new Java version. Therefore, nine combined local and line search versions could be chosen by the users. Now the algorithms and the efficiency of these new variants will be discussed as follows.

2.2.1 Unirandi

The Unirandi [16] local search method was originally part of GLOBAL. We have updated it so that arbitrary line search technique can be attached. In this way, we created UnirandiCLS, Unirandi with Custom Line Search. The pesudocode can be followed in Algorithm 1.

Unirandi performs line search along randomly generated directions. A trial point is computed based on the current point, on the actual generated direction, and on a step length parameter. If the current point fails to reduce the best function value the negative direction will be tried. The algorithm decreases the step length parameter after two consecutive failures along two generated directions. In a successful case (the actual function value is smaller than the best one), an arbitrary line search technique can be attached which performs further function reductions.

Algorithm 1 Unirandi local search method with custom line search

Step 1. While the maximal number of function evaluations is not reached or the change in function values or vectors are larger than a threshold value do:
Step 2. Generate a random direction.
Step 3. Run a line search algorithm.
Step 4. If line search succeeded go to Step 1.
Step 5. Turn towards the opposite direction.
Step 6. Run a line search algorithm.
Step 7. If line search succeeded go to Step 1.
Step 8. Increase the step length and check the number of iterations, if it is less than 2 go to Step 1.
Step 9. Decrease the step length and set the number of iterations to zero.
Step 10. End while.
Step 11. Save new optimum point.

2.2.2 Rosenbrock

The Rosenbrock [17] method achieves nonlimited search directions by rotating the axes, but still searching along them. The essence of the method is that it rotates one of the axes towards the most favorable direction and it continues searching along the rotated coordinate system's axes.

After a successful step, it increases the step length. After an unsuccessful step, it decreases the step length and turns in the opposite direction. This process is continued until at least one function evaluation becomes successful in each coordinate direction. After this, the axes are rotated again. The stopping criteria are examined after each transformation.

The Rosenbrock method is rotating the axes towards the best vector using the Gram-Schmidt orthogonalization process. The best vector is determined by the sum of the starting vector and the best vector.

Algorithm 2 Rosenbrock local search method with custom line search

Step 1. While the maximal number of function evaluations is not reached or the change in function values or vectors are larger than a threshold value do:
Step 2. For: iterate through every dimension.
Step 3. Do:
Step 4. Run a line search algorithm.
Step 5. If the line search is not successful: turn towards the opposite direction.
Step 6. While line search is not successful and maximum one direction is tested at the same time.
Step 7. If we don't have an unsuccessful step in any coordinate direction: rotate coordinates.
Step 8. Else: halve step length.
Step 9. End for.
Step 10. End while.
Step 11. Save the new optimum point.

2.2.3 NUnirandi

A disadvantage of the Unirandi local search method is that it is not effective on ill-conditioned problems. These problems can be characterized by long, and almost parallel contour lines, hence function reduction can only be achieved along a few directions. By searching along in random directions Unirandi has difficulty finding good directions, especially on high dimensional problems. Many problems, especially real-life ones have an ill-conditioned nature, so it is beneficial to implement methods that can cope with this kind of problem.

Both Rosenbrock and NUnirandi [18] (New Unirandi) use the advantages of random directions, which idea comes from Unirandi. But NUnirandi, just like

Rosenbrock, follows the direction in which we will hopefully find the optimum. Once we have found a good direction, we do a few more function evaluations there [12].

The algorithm is mostly the same as Unirandi, described earlier, but with an effective modification in the algorithm. Namely, the method tries to make further line searches along the last two saved pattern directions.

Algorithm 3 NUnirandi local search method with custom line search, supplement to the original Unirandi

Step 10 1/2-a. For: iterate through the last two saved pattern directions:
Step 10 1/2-b. Evaluate the new point based on the best point, step length and direction.
Step 10 1/2-c. Run a line search algorithm.
Step 10 1/2-d. If the line search succeeded go to Step 10 1/2-a.
Step 10 1/2-e. Turn towards the opposite direction.
Step 10 1/2-f. Run a line search algorithm.
Step 10 1/2-g. If the line search succeeded go to Step 10 1/2-a.
Step 10 1/2-h. End for.

2.3 Line Search

An important component of most local search methods is the line search technique. Implementation of the local search method is crucial in terms of function evaluations. Hence, carefully designed line search algorithms are welcomed.

As the line search is a separate module in the new Java implementation, it can be attached easily to the local search method. A line search algorithm should receive a decent direction and a starting point, and in the end, it returns a point with a corresponding function value.

2.3.1 Doubling Stepper

The algorithm you can see below was originally a part of Unirandi, thus we isolated and developed it further. The method moves as far as possible in the search direction until the function stops decreasing. Fast progress is ensured by the duplication of the step length after each successful step -- this is where the method got its name from.

Algorithm 4 Doubling stepper line search method

Step 1. Move by step length towards the search direction and evaluate the new point.
Step 2. While the new function value is smaller than the previous one do:
Step 3. Double step length.
Step 4. Move by step length towards the search direction and evaluate the new point.
Step 5. End while.

2.3.2 Function Fit

We can use the results we obtained from the doubling stepper's unsuccessful steps as base points and fit a curve on them. By fitting a curve near a possible optimum, the minimum point of the curve and the optimum we are searching for can be close. In this way in the case of some functions-which have sections where we can fit a quadratic or biquadratic curve-we can find the optimum way faster.

Algorithm 5 Function fit line search method that fits a curve on three/five points, extension after doubling stepper
Step 5 1/2-a. If: we have enough (three/five) control points:
Step 5 1/2-b. For: iterate through the dimensions:
Step 5 1/2-c. Fit a curve and evaluate its minimum point.
Step 5 1/2-d. End for.
Step 5 1/2-e. Save the last three/five best points.
Step 5 1/2-f. End if.

Fitting on three starting points. If the doubling stepper made at least three unsuccessful steps, it means that we have at least three starting points, where we know the function values. Then, we can fit a quadratic curve, and we can obtain the minimum point of this curve by using the formula:

$$x = b - \frac{\frac{1}{2}\left((b-a)^2(f_b-f_c)-(b-c)^2(f_b-f_a)\right)}{(b-a)(f_b-f_c)-(b-c)(f_b-f_a)} \quad (1)$$

If the minimum value we get with the formula is better than the previously-stored optimum, then we use the new one for further calculations.

Fitting on five starting points. When we are fitting on five starting points, we need some preprocessing before calculating the minimum value of the function's minimum point. First, we remove the duplicated points, then we create the matrix A for elimination. We use Gaussian elimination with partial pivoting. Afterward, we can get the minimum value depending on the number of coefficients using the cubic equation solution (Cardano-formula), quadratic formula, or just simply divide variable c with variable a.

3 Results

Previously, GLOBAL has been compared with common optimization procedures on standard test functions. In these comparisons, the standard Unirandi and NUnirandi were used in GLOBAL with a simple doubling stepper line search [19, 20]. In this work as well the standard test functions were used, e.g. Branin, Goldstein Price, Six-Hump Camel, Zakharov and Hartmann in several dimensions [15]. The directional choice of the local search method was combined with the three-line search methods so that a total of nine different algorithms were tested on

more than 60 standard test functions. Using the stopping criterion used by László Pál et al [11], we were able to compare our results and illustrate the improvement. Therefore, we specified that the number of function evaluations in the global search should not exceed 100.000, and the sample size was set to fifty and the alpha parameter to 0.1. The algorithms were run until they reached a value correct to six decimal places, or 100.000 function evaluations, and we analyzed how many function evaluations (FE) were needed. During the computation, each test function was executed one hundred times and its average was calculated. The success rate (SR) shows how many times the global optimum was reached. Finally, the line search we developed was compared to the original one and shows the improvement in function evaluation (%) between the original line search method and the new one.

Improvement is defined by dividing the results we get from the original method (the one obtained by the doubling stepper) with the results coming from the function fitting. For this reason, when reviewing *Table 1* we need to look for the rows where the percentages are below zero. In these cases, we needed less function evaluations with the new method. The local Rosenbrock search method with the five-point-based fitting line search works better in 78% of solved test cases. For the functions Cigar, Sphere, and Sum Squares, we were able to decimate the number of evaluations because these have sections where their curves are a quadratic function, so that the fitting can find the local optima almost perfectly when the line search direction is correct.

The results show that the algorithm is efficient with this type of function, and if the local optimal environment is different, its efficiency is worse. The Rosenbrock method with the three-point-based fitting allows us to achieve an improvement in 60% of the solved test cases.

Table 1

The Success Rate (SR) and the number of Function Evaluations (FE) using the Rosenbrock local search method on different functions in several dimensions with the percentage of improvement (%) using the various line search techniques

Function		Rosenbrock		Fit3			Fit5		
	dim	SR.	FE.	SR.	FE.	%	SR.	FE.	%
Booth	2	1.00	321	1.00	317	-1	1.00	274	-15
Cigar-40	40	0.23	77772	0.17	71733	-8	1.00	10286	-87
Cigar-5	5	1.00	1592	1.00	1529	-4	1.00	946	-41
Discuss-40	40	1.00	40753	1.00	40759	0	1.00	40597	0
Discuss-5	5	1.00	3801	1.00	3510	-8	1.00	3679	-3
Griewank-5	5	0.30	59450	0.23	58292	-2	0.16	57722	-3
Hartman-3	3	1.00	699	1.00	693	-1	1.00	588	-16
Hartman-6	6	1.00	2549	1.00	2457	-4	1.00	2090	-18
Levy	5	0.43	21142	0.40	17374	-18	0.42	17543	-17
Matyas	2	1.00	375	1.00	360	-4	1.00	289	-23
Perm-(4.1/2)	4	0.22	59960	0.26	54334	-9	0.34	48054	-20

Perm-(4.10)	4	0.30	10670	0.18	11983	12	0.29	10301	-3
Power sum	4	0.03	75427	0.06	54718	-27	0.12	53936	-28
Rastrigin	4	0.02	29074	0.04	22214	-24	0.04	25507	-12
Rosenbrock-5	5	1.00	4780	1.00	5687	19	0.99	4702	-2
Schaffer	2	0.47	45411	0.42	42124	-7	0.52	49749	10
Shekel-10	4	0.93	25034	0.93	18441	-26	0.96	17465	-30
Shekel-5	4	1.00	5346	1.00	6133	15	1.00	5627	5
Shekel-7	4	0.96	16917	0.95	17112	1	0.96	24148	43
Shubert	2	1.00	475	1.00	523	10	1.00	457	-4
Six hump	2	1.00	241	1.00	247	2	1.00	226	-6
Sphere-40	40	1.00	10002	1.00	9870	-1	1.00	3516	-65
Sphere-5	5	1.00	733	1.00	706	-4	1.00	380	-48
Sum squares-40	40	1.00	43011	1.00	42761	-1	1.00	24532	-43
Sum squares-5	5	1.00	816	1.00	836	2	1.00	860	5
Zakharov-5	5	1.00	1035	1.00	1045	1	1.00	986	-5
Zakharov-40	40	1.00	46244	1.00	46487	1	1.00	45293	-2

Since Unirandi does not take advantage of the calculation and use of the favorable direction, this method has not improved as much as the others (see Table 2). When using the five-point-based fitting method, the results are usually a few percent better. Although we have seen a significant improvement with the fitting, there are features where it performs noticeably worse. In 65% and 48% of the solved test cases, the two methods were an improvement.

In all cases, the success rate is mainly not changed. A line search substantially affects the number of function evaluations only in the right direction.

Table 2

The Success Rate (SR) and the number of Function Evaluations (FE) using the Unirandi local search method on different functions in several dimensions with the percentage of improvement (%) using the various line search techniques

Function		Unirandi		Fit3			Fit5		
	dim	SR.	FE.	SR.	FE.	%	SR.	FE.	%
Booth	2	1.00	299	1.00	301	1	1.00	269	-10
Cigar-40	40	0.00	0	0.00	0		0.00	0	
Cigar-5	5	0.24	90044	1.00	73963	-18	1.00	87779	-3
Discuss-40	40	1.00	42550	1.00	32731	-23	1.00	36797	-14
Discuss-5	5	1.00	9774	1.00	7614	-22	1.00	8415	-14
Griewank-5	5	0.38	57091	0.37	56398	-1	0.40	57067	0
Hartman-3	3	1.00	995	1.00	811	-19	1.00	945	-5
Hartman-6	6	1.00	4484	1.00	3979	-11	1.00	5361	20
Levy	5	0.46	18070	0.58	20163	12	0.42	21427	19
Matyas	2	1.00	335	1.00	343	3	1.00	318	-5
Perm-(4.1/2)	4	0.01	23111	0.04	22644	-2	0.03	44854	94
Perm-(4.10)	4	0.00	0	0.00	0		0.00	0	
Power sum	4	0.04	34460	0.05	27285	-21	0.04	16656	-52

Rastrigin	4	0.00	0	0.00	0		0.02	43683	
Rosenbrock-5	5	0.01	79881	0.02	98684	24	0.00	0	
Schaffer	2	0.67	35937	0.61	39623	10	0.54	45265	26
Shekel-10	4	0.98	19713	0.96	15303	-22	0.98	16026	-19
Shekel-5	4	1.00	5276	0.99	7180	36	1.00	5781	10
Shekel-7	4	0.98	21659	0.96	17313	-20	0.96	18538	-14
Shubert	2	1.00	393	1.00	384	-2	1.00	387	-2
Six hump	2	1.00	170	1.00	167	-2	1.00	171	1
Sphere-40	40	1.00	4813	1.00	4833	0	1.00	5471	14
Sphere-5	5	1.00	556	1.00	548	-1	1.00	573	3
Sum squares-40	40	1.00	35566	1.00	33610	-5	1.00	40857	15
Sum squares-5	5	1.00	689	1.00	650	-6	1.00	714	4
Zakharov-5	5	1.00	783	1.00	781	0	1.00	857	9
Zakharov-40	40	1.00	27081	1.00	27313	1	1.00	30281	12

As far as the efficiency increase of NUnirandi is concerned, NUnirandi executed at least 65% of the solved test functions with fewer function evaluations with both line search methods (See Table 3). All test functions achieved fewer evaluations than Rosenbrock. And just as with Rosenbrock, we get more stable results by adjusting to five points.

The performance graphs (see Figure 1) show the percentage of tasks that found the global optimum in a given number of function evaluations. As you can see, the function of fit3 and fit5 is generally above the traditional doubling step in all cases. That is, in addition to a similar number of evaluations, it has already found the global optimum in several cases. Since it is never below the traditional doubling step in terms of test cases, its use should not be harmful in all cases.

Table 3

The Success Rate (SR) and the number of Function Evaluations (FE) using the NUnirandi local search method on different functions in several dimensions with the percentage of improvement (%) using the various line search techniques

Function		NUnirandi		Fit3			Fit5		
	dim	SR.	FE.	SR.	FE.	%	SR.	FE.	%
Booth	2	1.00	332	1.00	334	0	1.00	291	-12
Cigar-40	40	1.00	32283	1.00	30780	-5	1.00	18673	-42
Cigar-5	5	1.00	1503	1.00	1557	4	1.00	1276	-15
Discuss-40	40	1.00	36503	1.00	28495	-22	1.00	32039	-12
Discuss-5	5	1.00	6117	1.00	5110	-16	1.00	5342	-13
Griewank-5	5	0.38	47461	0.37	50819	7	0.46	48210	2
Hartman-3	3	1.00	352	1.00	339	-4	1.00	334	-5
Hartman-6	6	1.00	1141	1.00	1437	26	1.00	1449	27
Levy	5	0.39	16631	0.50	23831	43	0.50	21598	30
Matyas	2	1.00	348	1.00	358	3	1.00	323	-7
Perm-(4.1/2)	4	0.71	44912	0.73	42684	-5	0.79	37474	-17
Perm-(4.10)	4	0.30	8157	0.25	7969	-2	0.26	7469	-8

Power sum	4	0.64	51886	0.77	47936	-8	0.98	39377	-24
Rastrigin	4	0.01	60261	0.01	62353	3	0.05	35100	
Rosenbrock-5	5	0.99	5284	1.00	5093	-4	0.99	4739	-10
Schaffer	2	0.62	38196	0.54	34742	-9	0.61	35514	-7
Shekel-10	4	0.95	16823	0.96	21124	26	0.98	17190	2
Shekel-5	4	1.00	6200	1.00	4685	-24	1.00	6348	2
Shekel-7	4	0.96	17637	0.97	15246	-14	0.95	21256	21
Shubert	2	1.00	470	1.00	425	-10	1.00	456	-3
Six hump	2	1.00	198	1.00	197	0	1.00	178	-10
Sphere-40	40	1.00	4949	1.00	4907	-1	1.00	5576	13
Sphere-5	5	1.00	608	1.00	606	0	1.00	633	4
Sum squares-40	40	1.00	16198	1.00	16098	-1	1.00	14968	-8
Sum squares-5	5	1.00	714	1.00	682	-4	1.00	730	2
Zakharov-5	5	1.00	813	1.00	828	2	1.00	869	7
Zakharov-40	40	1.00	24075	1.00	24100	0	1.00	26702	11

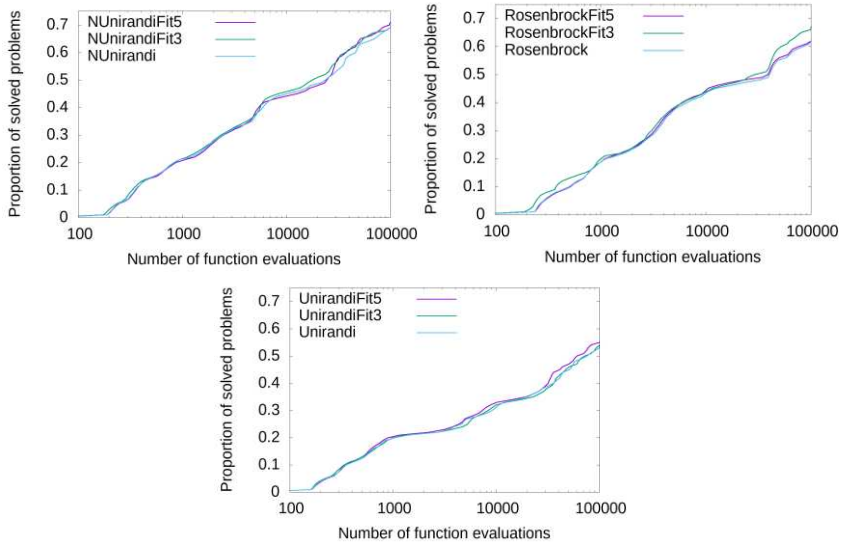


Figure 1
Proportion of the solved problems

Unfortunately, we could not achieve an improvement by adjusting more curves one after the other. So we won't make a further discussion about this task.

Conclusions

We can see that for those functions where the right direction can be found, a quite satisfactory improvement in local search can be achieved. We can therefore say that it is worth using our method and that we have achieved a certain improvement on more than 50-70% of our test functions. This result is not prominent on the whole

test set, as can be seen in the figures. But it is very effective for objective functions with a quadratic-like local optimum. Also, in the other test cases, the number of evaluations will not be much higher.

In summary, the modularization implementation allows us to quickly and effectively test huge amounts of test functions with different combinations. Respectively, with all the new line search versions, the user is not expected to perform worse than the previous version. Unfortunately, there is no clear choice between the new versions, but this modularzation allows us to find the best optimizer for each test function.

Acknowledgement

This research was supported by the projects "Extending the activities of the HU-MATHS-IN Hungarian Industrial and Innovation Mathematical Service Network" EFOP-3.6.2-16-2017-00015, the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, and the Unkp-19-4-Bolyai+ New National Excellence Program of the Ministry of Human Capacities.

References

[1] Banga, J. R., C. G. Moles, and A. A. Alonso, Global Optimization of Bioprocesses using Stochastic and Hybrid Methods, Frontiers in Global Optimization, 45-70 (2003)

[2] Moles, C. G., J. R. Banga, and K. Keller, Solving nonconvex climate control problems: pitfalls and algorithm performances, Applied Soft Computing 5, 35-44 (2004)

[3] Balogh, J., T. Csendes, and R. P. Stateva, Application of a stochastic method to the solution of the phase stability problem: cubic equations of state, Fluid Phase Equilibria 212, 257-267 (2003)

[4] Balogh, J., T. Csendes, and T. Rapcsák, Some Global Optimization Problems on Stiefel Manifolds, J. of Global Optimization 30, 91-101 (2004)

[5] Balogh, J., T. Csendes, and R. P. Stateva, Application of a stochastic method to the solution of the phase stability problem: cubic equations of state, Optimization Letters 2, 445-454 (2008)

[5] Csendes, T., Nonlinear parameter estimation by global optimization efficiency and reliability, Acta Cybernetica 8, 361-370 (1988)

[6] Bánhelyi, B., T. Csendes, and B. M. Garay, A veried optimization technique to bound topological entropy rigorously, Proceedings of the SCAN-2006 Conference, IEEE (2007)

[7] Csete, M., G. Szekeres, B. Banhelyi, A. Szenes, T. Csendes, and G. Szabo, Optimization of Plasmonic structure integrated single-photon detector designs to enhance absorptance, Advanced Photonics 2015, JM3A.30 290 (2015)

[8] Bánhelyi, B., T. Csendes, B. M. Garay, Optimization and the Miranda approach in detecting horseshoe-type chaos by computer, Int. J. Bifurcation 270 Chaos 17, 735 (2007)

[9] Csendes, T., B. Bánhelyi, and L. Hatvani, Towards a computer-assisted proof for $\sum^3$ chaos in a forced damped pendulum equation, J. Computational and Applied Mathematics 199, 378-383 (2007)

[10] Csendes, T., B. M. Garay, and B. Bánhelyi, A veried optimization technique to locate chaotic regions of Hénon systems, J. of Global Optimization 35, 145-160 (2006)

[11] Csendes, T., L. Pál, J. O. H. Sendín, J. R. Banga, The GLOBAL Optimization Method Revisited, Optimization Letters 2, 445-454 (2008)

[12] Pál, L., An Improved Stochastic Local Search Method in a Multistart Framework, Proceedings of the 10th Jubilee IEEE International Symposium on Applied Computational Intelligence and Informatics, 1170 (2015)

[13] Sendín, J. O. H., J. R. Banga, and T. Csendes, Extensions of a Multistart Clustering Algorithm for Constrained Global Optimization Problems, Industrial & Engineering Chemistry Research 48, 3014-3023 (2009)

[15] Bánhelyi, B., T. Csendes, B. L. Lévai, L. Pál, and D. Zombori, The updated GLOBAL optimization algorith: Newly Updated with Java Implementation and Parallelization, book, SpringerBriefs in Optimization, Springer (2019)

[16] Järvi, T., A random search optimizer with an application to a max-min problem, Publications of the Institute for Applied Mathematics, University of Turku, No. 3 (1973)

[17] Rosenbrock, H. H., An Automatic Method for Finding the Greatest or Least Value of a Function, Computer J. 3, 175-184 (1960)

[18] Pál, L., Empirical study of the improved UNIRANDI local search method, CEJOR, Vol. 25, 4: 929{952 (2017)

[19] Pál, L., T. Csendes, MC. Markót, A Neumaier, Black box optimization benchmarking of the global method, Evolutionary computation 20 (4), 609-639, (2012)

[20] Pošík P., W. Huyer, L. Pál, A comparison of global search algorithms for continuous black box optimization, Evolutionary computation 20 (4), 509-541 (2012)

Fake News Detection Related to the COVID-19 in Slovak Language Using Deep Learning Methods

Martin Sarnovský, Viera Maslej-Krešňáková, Klaudia Ivancová

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 04001 Košice, Slovakia. E-mail: martin.sarnovsky@tuke.sk, viera.maslej.kresnakova@tuke.sk, klaudia.ivancova@student.tuke.sk

Abstract: One of the biggest problems nowadays in the online environment is the spreading of misinformation. Especially during a global pandemic, the most popular topics of fake news are related to coronavirus. Therefore, automatic detection of such news in the online media or social networks can help with the prevention of misinformation spreading. During the recent years, deep learning models proved to be very efficient in this task. However, the majority of the research focuses on the training of these models using publicly available data collections, mostly containing news articles written in the English language. As the spreading of fake news is a global phenomenon, it is also necessary to explore these approaches on the various local data sources. The work presented in this paper focuses on using the deep learning models for the automatic detection of fake news written in the Slovak language. We collected the data from multiple local online news sources related to the COVID-19 pandemic and used it to train and evaluate the various deep learning models. Thanks to the combination of bidirectional long-short-term memory network with one-dimensional convolutional layers, we achieved an average macro F1 score on an independent test set of 94%.

Keywords: natural language processing; deep learning; convolutional neural networks; fake news; COVID-19

1 Introduction

In recent years, people are spending more and more of their lives online and on social media. There are many advantages and disadvantages in moving human activity and communication to the online environment. The ability to read, share and publish information for all equally is considered the most significant benefit. Exchanging information via the Internet will take much less time, money, and effort. Despite the fact that the quality of news appearing on social media is much lower than in traditional and verified sources, thanks to the mentioned advantages,

people tend to search for such news more and more. While the vast majority of content on the Internet is user-generated content, the quality of the content varies. The majority of the users try to produce true and decent content; however, there are also users who create misleading content in the online environment.

Authors in [1] define anti-social behavior as behavior that violates the fundamental rights of others. It is characterized by repeated violations of social rules, defiance of authority, and the rights of others. There are several types of anti-social behavior in the online environment. One specific type is related to the production and spreading of hoaxes, fake news, and false reviews. They are purposefully created for the dissemination of misinformation, concealment of true information, misleading the reader, or false beliefs of readers. Their occurrence tends to increase during public events, disasters, or when evaluating new products. In the work presented in this paper, we focused on the detection of fake news. Fake news can be defined [2] as news articles that are intentionally and demonstrably untrue and misleading their readers. There are two basic properties of fake news related to this definition, namely the authenticity of the message, according to which we can retrospectively verify false information and the intention under which fake news is created. As already mentioned, most often, such reports are created for misleading consumers and generally for dishonest intent. The authors in [3] supplemented this definition with the following terms, which they do not consider to be false in this respect:

- Satirical news - contain appropriate content that is not intended to mislead or mislead consumers
- Rumors - which do not come from news
- Conspiracy theories - which are sometimes difficult to describe as real or false
- Misinformation - which is created unintentionally
- Hoaxes - which are created to entertain or deceive an individual

Fake news is not a new problem, as the type of traditional media has changed from print newspapers to radio and television to online news, so has fake news. However, social media and the increasing use and living through the online environment have helped make this problem a major one. From a psychological point of view, spreading misinformation is very easy because people naturally cannot distinguish very well between true and false news. Traditional fake news is aimed mainly at consumers, where their vulnerability factors can be taken into account. Also, many users spread fake news and trust them to gain power or acceptance by others in the community or even to satisfy societal views and needs. Fake news is usually spreading on social media through specific accounts. Due to their low cost and rapid dissemination of messages, they are widely used today. These include social robots, trolls, and cyborgs. Such social robots are created to share content and interact with people in the online environment, primarily using

computer algorithms. They are designed specifically for the purpose of manipulating and spreading false messages. For example, during the 2016 US presidential election, more than 19% of online conversation were found to have been manipulated presidential election results information in terms of spreading false news and disrupting online communities [4].

The solution how to prevent the massive spread of anti-social behavior and false news in the online environment involves the creation of tools for early detection and elimination of such information. Today, manual techniques of detection of such news are being employed (e.g., human moderators, etc.), but they are insufficient as the number of information increases. To eliminate the impact of fake news that would achieve the desired results, it is necessary to create automated tools for their detection. More recently, such approaches can be helpful more than ever, especially during the COVID-19 pandemic. Misinformation related to the pandemic and vaccination are spreading through social media rapidly, and similar tools able to detect them may be helpful in the prevention of their reach.

Moreover, as the pandemic is a global phenomenon, much of the misinformation spreads locally, in different countries, via various local web portals. Therefore, it may be interesting to train the models using the data in particular languages. In this paper, we focused on training the deep neural network models able to detect fake news from the news articles related to COVID-19 written in the Slovak language. We decided to use the different architectures of deep learning models, as they proved very efficient when applied to the related problems (e.g., fake news detection in various domains and languages). In our research, we focused on using only the textual content of the articles to capture the linguistic features that distinguish the regular articles from the misinformation.

The presented paper is organized as follows: Section 2 is dedicated to the description of existing approaches using neural networks for fake news detection. The following section summarizes the data collection and pre-processing steps. Then, in Section 4, we describe models training and evaluation of the experiments.

2 Related Work

Fake news detection is usually considered a binary classification task, where the models predict, based on the content, if a particular news piece contains fake news or not. Neural networks are among the most frequently used methods in the area of automatic detection of fake news from text. Besides the modeling, many works focus on the text pre-processing and appropriate representation. For example, authors in [3] claim that when working with traditional news sources, it is sufficient to work only with the content of the news piece. On the other hand,

when working with the social media posts (or discussion forums or similar sources), information related to the source, attachments (e.g., pictures, videos) may be useful as well. The title of the news article is usually important, as fake news often uses strong or outrageous content, so-called *clickbait*, which forces the user to click on the article and read it. Therefore, it is necessary to explore the linguistic aspects and, besides them, also explore the information related to the authors of the news, including reactions of the readers [5]. Authors in [6] compared neural network models trained using full texts from the articles and just the title text. When comparing the evaluation metrics on the full-text data to title texts, the models still managed to perform on a similar level. One of the reasons may be using of clickbait in the title texts.

Different network architectures have been used to tackle the automatic fake news detection from the text. Convolutional neural networks were used in [7] for the detection of fake news in texts containing political statements. Authors also used the metadata, describing the authors' info (e.g., occupation) or information related to other author's statements. Authors randomly initialized an embedding vector matrix to encode the data and metadata and used the convolutional layer of the neural network to capture the dependence between the metadata vectors. Next, he performed a maximum association operation in the latent space, followed by the LSTM layer of the recurrent neural network. Finally, the author combined the representations of the texts with the metadata representation from LSTM and brought them into a fully connected layer to generate the final prediction. The Word2vec tool was used to create the embedding.

In [8], the authors also aimed to detect fake news using the Capsule neural network. They used this model to improve classical CNN and RNN by adding specific properties to each source and destination node. The model created by the authors is used to identify fake news articles with different lengths. Depending on the size of the pieces, the authors used two different architectures. The model uses pre-trained vectors to initialize learning. For the short texts, the authors developed a structure whose layers are identical to those in the first model, but only two parallel networks are considered. In this model, static word embedding is used, which represents pre-trained vectors that are kept static during training, and only other parameters are trained. The model containing medium and long texts achieved the best accuracy using a non-static word embedding 99.8%. The model containing short texts was still evaluated using metadata because it was more difficult to detect false messages.

In [9], the authors used a dataset containing reports from the period of the US presidential election in 2016. They used a deep neural network as a model and solved the problem of binary classification. The first layer in the neural network consists of pre-trained word embedding using Word2vec. Embedding is used as an input to a convolution layer with 128 filters and a window size of 3. For evaluation, the authors used an accuracy metric where they were able to achieve 93.5% accuracy. The authors in [10] also used content from the US

presidential election in 2016 - toxic comments from Twitter. They provide an overview of various pre-processing options, standard deep learning models, and popular transformers models.

In [11], the authors worked on designing a deep convolutional neural network to detect fake news. They developed the model so that the functions learn to automatically distinguish the elements for classifying fake news through several hidden layers built into the neural network. The authors used the uncontrolled GloVe learning algorithm as an embedding model. The model was followed by three parallel convolution layers, a maximum common layer, and finally, an output layer based on prediction. They also used a single flatten layer, which converts elements taken from a common layer and maps them to a separate column, which is then moved to a fully connected layer. The authors used ReLU as the activation function, the primary function of which is to remove negative values from the activation map by setting it to zero. By evaluating the model, they managed to obtain an accuracy of 98.36%.

When considering the detection of fake news in the Slovak language, the work [12] is the only one that explores the dataset of texts in the Slovak language. The aim was to explore different approaches to detecting fake news based on morphological analysis. The authors have created their own data set, which contains articles in the Slovak language from the local news sources. The authors used articles containing the keywords "NATO" and "Russia". These were classified into two specific classes according to the publisher's source using the web portal *www.konspiratori.sk*. Since the Slovak language has complex rules for declension, the authors have decided that the use of morphologically annotated corpora from the Slovak National Corpus will contribute to automated morphological analysis. The morphological analysis was applied to all articles in the dataset, and each word was assigned a set of morphological tags. Contrary to other works, the authors did not use the neural network but used the decision tree model. They divided the analyzed data set in the ratio 70:30 and set the different depths of the tree with the model using entropy. With the decision tree model, they achieved an accuracy of 75%.

3 Data Collection and Pre-Processing

In the work presented in this paper, we focused on detecting fake news in Slovak online space. To obtain the data from various local online newspapers, we used the MonAnt platform [13]. We used the platform to collect the news articles related to COVID-19 pandemic from mainstream local newspapers, as well as from unreliable sources, often publishing conspiratory content. In general, we focused on covering different types of sources to be able to collect both regular news articles as well as misleading pieces. In the MonAnt platform, we created

connectors to the web news portals *Aktuality, Hospodárske noviny, TA3, Hnonline, Slobodný Vysielač, Zem & Vek, Magazín pán občan, Hlavné správy, Proti prúdu, Rady nad zlato* etc. and filtered the articles containing the selected keywords: *Covid, Covid-19, Coronavirus, SARS-CoV-2.*

To train the models using such data, we needed to obtain the class label. At first, the target feature was derived according to the *www.konspiratori.sk* database. The database is a result of the local media experts aiming to monitor and reporting of fake news in different local media and contains a ranking of many regional online news portals and their respective "score", representing the probability of publishing misleading information and fake news in their sites. We used the score for initial labeling and derived the binary target feature according to the trustworthiness of the sources, separating the reliable sources (with very low scores) from the suspicious ones (the highest scores). However, such labeling only considered the credibility of the whole source (e.g., website or newspaper) but not the credibility of individual reports. Such an approach may lead to incorrect labeling, as many conspiratory websites also publish regular news. We've concluded that the best way to assign relevance to articles correctly is to manually re-label the content from the conspiratory websites and correct the labels for the articles, which contain standard information (e.g., re-published news from the press agencies, etc.). After such correction, the dataset consisted of 12,885 documents containing regular news and 851 articles labeled as fake news. The target attribute was heavily imbalanced. Usually, in such cases, application of some of the balancing techniques (e.g. over/undersampling, or more advanced SMOTE) is very common. However, after conduction of preliminary experiments on a similar data [14] we found out, that such modification is not necessary in this case.

During the pre-processing phase, we removed all punctuation marks, non-alphabetic characters, hypertext links not essential for the detection of fake news and kept only the letters of the Slovak alphabet. We also removed words that referred to the article's source (e.g., writing or photo credits). We wanted to focus solely on the textual content of the news piece; such data in the text may present an information leak about the source; therefore, we decided to remove them. Besides that, also stopwords were removed (e.g., prepositions, conjunctions, pronouns, etc.).

To train the models, it was necessary to convert the text content of the articles into a vector representation. Vector word representation or word embeddings is a technique where individual words are represented as vectors with a real value in a predefined n-dimensional vector space. Word embeddings capture the semantic and syntactic meaning, so almost similar vectors represent similar words placed close together in vector space. The result of word embeddings is a coordinate system in which similar words are placed close together. In our work, we used Word2vec embeddings. Word2vec is a technique for natural language processing published in 2013 [15]. The Word2vec algorithm uses a neural network model to learn word associations from a large corpus of text.

4 Models Training and Evaluation

In the experiments, we used the convolutional neural network (CNN), Long-Short-Term Memory (LSTM) network, and a CNN combined with the bidirectional LSTM. CNN is one of the most popular types of deep neural networks. The main advantage of CNN is that it automatically detects the important features without any human supervision. This is why CNN would be an ideal solution to computer vision and image classification problems. The benefit of using CNNs for sequence classification is that they can learn from the raw time series data directly, and do not require domain expertise to manually engineer input features. CNN's use the convolution operation instead of the general multiplication of matrices in at least one of the layers in their architecture. One dimensional convolutional layer creates a convolution kernel that is convolved with the layer input over a single spatial (or temporal) dimension to produce a tensor of outputs. A typical CNN architecture consists of three layers, a convolution layer, a pooling layer, and a fully connected layer. Layers are used to analyze images, objects, speech, or language features. LSTM is a recurrent neural network introduced by [16]. It is modified to solve the vanishing gradient problem and can model sequences and long-term dependencies more accurately. The LSTM architecture contains special units called memory blocks located in a hidden layer instead of neurons. These blocks have memory cells with separate connections that remember and store the network's temporary state. They also contain multiplicative units that control the flow of information and are called gateways. Authors in [17] introduced a special type of recurrent neural network, namely the bidirectional recurrent neural network (BRNN). The idea is to bring each training sequence in both directions into two separate recurrent networks connected to the same output layer. Two independent RNNs create the BRNNs by dividing the state neurons into the part responsible for the forward and backward direction.

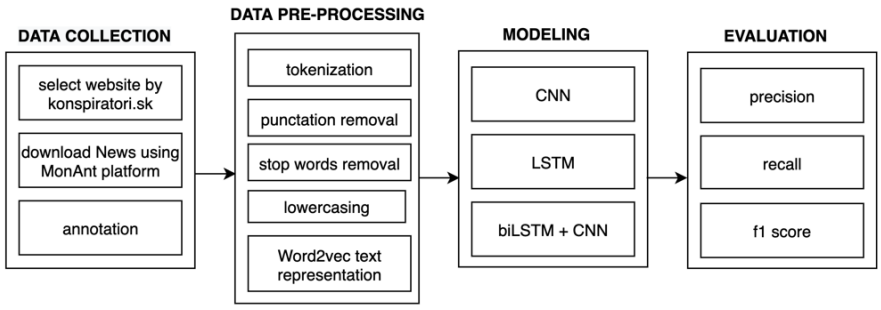


Figure 1
Workflow of the experiments

Figure 1 depicts the workflow of the experiments. The dataset consisted of 13,736 news articles, 12,885 regular, and 851 fake news. We divided the dataset into a training and test set. The training set contained 9,615 documents, of which 9,000 were relevant and 615 fake news. The test set, used for evaluating the model, contained 4,121 documents, of which 3,901 articles were regular news and 220 fake news. We used 10% of the training set to validate the model. The validation set was created by a stratified random split of the training set at each training of the model. Data pre-processing steps were described in Section 3.

4.1 The Architecture of the Models

In the phase of modeling, we used the following settings of models:

- activation function
    - hidden layers: ReLU
    - output layers: sigmoid
- loss function: binary cross-entropy
- optimizer: Adam
- regularization: Dropout, spatial dropouts

Because linear functions are severely limited and cannot recognize and learn complex patterns in the data needed to classify images, text, or sequences, neural network architectures use nonlinear activation functions [18]. Activation function is a function which decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias. We used Rectified Linear Unit (ReLU) - non-linear activation function used in nearly all modern neural network architectures. The output of ReLu is the maximum value between zero and the input value. Output is equal to zero when the input value is negative and the input value when the input is positive [19], [20]. The sigmoidal activation function is used as a nonlinear activation function on the output layer. We use it as the problem is a binary classification task, and usually, it can be used in conjunction with binary cross-entropy. This function transforms values in the range 0-1; thus, it determines the probability with which the input belongs to a given class.

In single-layer neural network architectures, the loss functions can be calculated directly from the weights. For training feedforward neural networks, error backpropagataion is used. Error backpropagation is about determining how changing the weights impact the overall loss in the neural network. The backpropagation works by computing the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time, iterating backward from the last layer [21], [22]. As a loss function, we used Binary Cross-Entropy (BCE). BCE is a type of loss function used in binary classification problems. The function is used in the neural network to predict the

probability of a given example that belongs to one of two classes. The activation function on the output layer is, in this case, a sigmoidal function.

Optimizers are algorithms used to change the attributes of neural networks (e.g., weights, learning rate) to reduce loss. In our work, we used the Adaptive moment estimation (Adam) optimizer. It is a method that calculates individual adaptive learning rates for each parameter. It is designed as a combination of Adagrad and RMSProp methods, where it takes advantage of both. Adagrad works well in sparse transitions, and RMSprop works well in online and non-stationary conditions. Both also maintain the speed of learning. The advantages of this optimizer are that it works with sparse gradients, is directly implementable, and does not require much memory. Overall, this model is considered robust and suitable for wide use in solving optimization problems in machine learning [23], [24].

Goodfellow [25] defined regularization as any adjustment made to a learning algorithm and aims to reduce generalization losses. The dropout regularization technique is one of the most used regularizations in the field of deep learning. It provides a computationally inexpensive but powerful method for a wide range of models. Dropout can prevent overfitting by temporarily removing neurons with all of their incoming and outgoing connections and forces a neural network to learn more robust features that are useful in conjunction with many different random subsets of the other neurons [26], [27]. Also, we used a spatial version of Dropout. This version performs the same function as Dropout; however, it drops entire 1D feature maps instead of individual elements. In this case, *SpatialDropout1D* will help promote independence between feature maps.

In experiments, we compared the following architectures:

- **CNN model**. The embedding layer was followed by a one-dimensional convolution layer with 100 filters and window size 2 and the activation function ReLu. 1D CNN was followed by the pooling layer - GlobalMaxPooling, whose output is input to a feedforward neural network with one dense layer with 256 neurons and Relu activation function. The output layer contained one neuron and a Sigmoid activation function.
- **LSTM model.** The embedding layer was followed in this model by LSTM layers with 128 units. In this experiment, we added a dropout regularization layer with a parameter of 0.2, representing 20% of the input neurons that will be deactivated with each epoch, thus preventing over-fitting. Then were continued one fully connected layer with 128 neurons, and ReLu activation function and an output layer with one neuron, and a sigmoid activation function.
- **biLSTM + CNN model.** The embedding layer was followed by spatial Dropout with parameter 0.2. The output from the regularization layer was the input to the bidirectional recurrent LSTM layer with 64 units, with a

recurrent dropout of 0.1. The architecture continues with a one-dimensional convolution layer that contained 32 filters and a window size of 3. It was followed by the GlobalMaxPooling layer, which represented the entrance to the feedforward neural network with one dense layer with 64 neurons and a ReLU activation function. Then, the dropout regularization layer with a parameter of 0.2 was used again, and the output layer contained one neuron and a sigmoid activation function. The architecture of the third model is shown in **Table 1**. The structure of this architecture was inspired by previous experiments in the classification of toxic comments [10].

Table 1
biLSTM + CNN model

Layer (type)	Output Shape	Parameters
Input Layer	(None, 1,000)	0
Embedding	(None, 1,000,100)	20,177,100
Spatial Dropout	(None, 1,000,100)	0
biLSTM	(None, 1,000,128)	84,489
Conv1D	(None, 998,32)	12,320
Global Max Pooling	(None, 32)	0
Dense	(None, 64)	2,112
Dropout	(None, 64)	0
Dense	(None, 1)	65
Total params.:	20,276,077	
Trainable params.:	20,276,077	
Non-trainable params.:	0	

4.2 Evaluation

We evaluated all trained models on an independent test set. To evaluation models we used following metrics:

- **Recall** = TP / (TP+FN),
- **Precision** = TP / (TP + FP),
- **F1 score** = 2 * (Precision * Recall) / (Precision + Recall),

where

- **TP – True Positive** examples are predicted to be fake news and are fake news;
- **TN – True Negative** examples are predicted to be relevant news and are relevant news;

- **FP – False Positive** examples are predicted to be fake news but are relevant news;
- **FN – False Negative** examples are predicted to be relevant news but are fake news.

Table 2
Evaluation of models on the test set

	Precision	Recall	F1 score	Support
CNN model				
Regular News	0.98	1.00	0.99	3,901
Fake News	0.98	0.62	0.76	220
Accuracy			0.98	4,121
Macro avg	0.98	0.81	0.88	4,121
Weighted avd	0.98	0.98	0.98	4,121
LSTM model				
Regular News	0.99	1.00	0.99	3,901
Fake News	0.96	0.78	0.86	220
Accuracy			0.99	4,121
Macro avg	0.97	0.89	0.93	4,121
Weighted avd	0.99	0.99	0.99	4,121
biLSTM + CNN model				
Regular News	0.99	1.00	0.99	3,901
Fake News	0.97	0.79	0.87	220
Accuracy			0.99	4,121
Macro avg	0.98	0.89	0.93	4,121
Weighted avd	0.99	0.99	0.99	4,121

Table 2 depicts the results of the experiments. In this task, we focused mainly on increasing the value of recall because we want to detect as much fake news as possible. The best accuracy with a value of 98.76% was achieved in the model with the third architecture. The second and third architectures achieve very similar results, but since the third model is more robust and the resulting recall value is one percent higher than in the LSTM architecture, we decided to continue working with biLSTM+CNN architectures.

In the modeling phase, we also performed the optimization of hyperparameters using the grid search method for the best-performing model. We used the following combination of the hyperparameters and their values:

- **Dropout rate –** 0.1, 0.2, 0.3
- **Batch size –** 16, 32, 64
- **Optimizer –** Adam, Stochastic gradient descent (SGD)

We obtained the best results using the dropout rate of 0.1, batch size 32, and Adam optimizer. We trained the best model with these settings and achieved an accuracy of 98.93 % on the test set. **Table 3** shows the overall results of the model after training the model using the best combination that came out in the grid search optimization with 5-fold cross-validation. **Table 4** depicts the confusion matrix of this model. We can observe that 34 regular news articles were classified as fake news while 10 fake news pieces were classified as regular.

Table 3
Evaluation of the CNN + biLSTM architecture after optimization

	Precision	Recall	F1 score	Support
CNN+biLSTM model after grid search with cross-validation				
Regular News	0.99	1.00	0.99	3,901
Fake News	0.95	0.82	0.89	220
Accuracy			0.99	4,121
Macro avg	0.97	0.92	0.94	4,121
Weighted avg	0.99	0.99	0.99	4,121

After optimization, the accuracy of the best model increased even more to 98.93%, which represents an increase compared to the previous best model by 0.17%.

Table 4
Confusion matrix of the best model

		Predicted values	
		Fake News	Relevant News
Actual values	Fake News	True Positive (186)	False Negative (34)
	Relevant News	False Positive (10)	True Negative (3991)

Conclusions

In the presented paper, we focused on the detection of fake news in the Slovak language using deep learning models. As most of the datasets contain news articles written in English, we had to collect the database of news pieces from various local online news portals. We decided to focus on the currently very popular and important topic and collected the news articles related to the COVID-

19 pandemic. Spreading misinformation, especially in this domain, can present a serious issue that may affect people's health and lives. Collected data were annotated using a combination of manual techniques and expert knowledge provided by a curated list of misinformation sources. In the experiments, we used deep learning models, which according to the literature, gains superior results in similar tasks. We used CNN, LSTM, and a combined CNN+biSLTM model, fine-tuned using grid search, which proved to perform with an accuracy of 98.93% and an F1 score of 94%. Although the results sound promising, they are heavily influenced by the data and annotation quality. In the future, we plan to extend the dataset for the training of such models (e.g., create more generic datasets, not just related to the pandemic) and improve the annotation quality. As the data volume grows, we plan to utilize the crowdsourcing approach to obtain the class labels for the data combined with active learning. Also, the initial data collection proved that the final dataset would be heavily imbalanced. In this area, it would be appropriate to deal also with augmentation techniques in future work, which would expand the dataset and increase the robustness of the models.

Acknowledgment

This work are supported by Slovak APVV research grant under contract No. APVV-16-0213 and Slovak VEGA research grant No. 1/0685/21.

References

[1] S. D. Calkins and S. P. Keane, “NIH Public Access,” Vol. 21, No. 4, pp. 1095-1109, 2010

[2] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *J. Econ. Perspect.*, Vol. 31, No. 2, pp. 211-236, 2017

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explor. Newsl.*, Vol. 19, No. 1, pp. 22-36, 2017

[4] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nat. Commun.*, 2019

[5] K. Machova *et al.*, “Addressing False Information and Abusive Language in Digital Space Using Intelligent Approaches,” in *Towards Digital Intelligence Society*, 2021, pp. 3-32

[6] V. M. Krešňáková, M. Sarnovský, and P. Butka, “Deep learning methods for Fake News detection,” in *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo)*, 2019, pp. 143-148

[7] W. Y. Wang, “‘ liar, liar pants on fire’: A new benchmark dataset for fake news detection,” *arXiv Prepr. arXiv1705.00648*, 2017

[8] M. H. Goldani, S. Momtazi, and R. Safabakhsh, “Detecting fake news with

capsule neural networks," *Appl. Soft Comput.*, Vol. 101, p. 106991, 2021

[9] N. O'Brien, S. Latessa, G. Evangelopoulos, and X. Boix, "The language of fake news: Opening the black-box of deep learning based detectors," 2018

[10] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, "Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification," *Appl. Sci.*, Vol. 10, No. 23, 2020

[11] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet – A deep convolutional neural network for fake news detection," *Cogn. Syst. Res.*, Vol. 61, pp. 32-44, 2020

[12] J. Kapusta, "Improvement of Misleading and Fake News Classification for Flective Languages by Morphological Group Analysis," 2020

[13] I. Srba *et al.*, "Monant : Universal and Extensible Platform for Monitoring , Detection and Mitigation of Antisocial Behaviour," 2019

[14] K. Ivancova, M. Sarnovsky, and V. Maslej-Kresnakova, "Fake news detection in Slovak language using deep learning techniques," in *SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 2021

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, Vol. 9, No. 8, pp. 1735-1780, 1997

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, Vol. 45, No. 11, pp. 2673-2681, 1997

[18] S. Sharma, "Activation functions in neural networks," *Towar. Data Sci.*, Vol. 6, 2017

[19] N. Ketkar and E. Santana, *Deep Learning with Python*, vol. 1. Springer, 2017

[20] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017

[21] R. Hecht-Nielsen, "Theory of the backpropagation neural network," 1989

[22] H. J. KELLEY, "Gradient Theory of Optimal Flight Paths," *ARS J.*, Vol. 30, No. 10, 1960

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014

[24] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 2018, pp. 1-2

[25] I. Goodfellow and Y. Bengio, “Deep Learning,” 2016

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0, 2012

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, Vol. 15, No. 1, pp. 1929-1958, 2014

Fixed Point, Iteration-based, Adaptive Controller Tuning, Using a Genetic Algorithm

István Lovas

Óbuda University - Doctoral School of Applied Informatics and Applied Mathematics, Bécsi út 96/b, H-1034 Budapest, Hungary
E-mail: lovas.istvan@nik.uni-obuda.hu

Abstract: From a control perspective, the exact conditioning of systems with time-varying parameters is still a challenge. Many adaptive control algorithms (Adaptive Inverse Dynamics – AID, Model Reference Adaptive Controllers – MRAC, etc.) exist today. "Fixed-Point Iteration Methods" attempt to offer "alternative" control planning methods to circumvent the application of the Lyapunov function technique. The foundations of the method were developed in 2009. RFPT is an iterative control method, based on the fixed-point theorem of Stefan Banach proved in 1922 [1]. There is usually no specific suggestion for the choice of controller parameters, as the response function also depends on the approximation model parameter used and the actual behavior of the system under control. Adaptive RFPT presupposes strongly nonlinear system models in the first place, so in this case, thinking in frequency image and step inputs is not relevant (it is not advisable to conflict a nonlinear system with step inputs), so it does not have a tuning technique applicable to LTI models. However, there are a number of optimal search methods that can also be used to tune controllers (e.g., PIDs), e.g. the Genetic Algorithm. Using this method, I developed a possible autotuning process for adaptive RFPT controllers.

Keywords: Adaptive Control; Fixed Point Iteration-based Adaptive Control; Banach Space; Genetic algorithm; GA-based RFPT auto-tuning; Auto-tuning method; Control systems

1 Robust Fixed-Point Transformation-based Adaptive Controller

From a control perspective, the exact conditioning of systems with time-varying parameters is still a challenge today. Many adaptive control algorithms (Adaptive Inverse Dynamics – AID, Model Reference Adaptive Controllers – MRAC, etc.) exist today, however, in many cases, their application is difficult, their fine-tuning is not trivial, and they demand significant mathematics knowledge, as in most cases the conditioning algorithm is based on the Lyapunov stability criterion.

Lyapunov in his Ph.D. dissertation defined several stability criteria. Their common characteristic is that they are physically based on a Lyapunov function, a differentiable function of time, that can be interpreted as a scalar error metric, which is not negative and is zero, exclusively in case of zero error, and this function has to be held in control (its time-based derivative must not be positive); this is ordinary stability. If it is accessible, that this derivative is negative enough so that it can make the Lyapunov function to converge to value zero in an infinite amount of time, an asymptotically stable system will be created [2] [3].

The biggest challenge is to select this function for a given specific control task; most of the time it is determined only by “intuitions”. The biggest problem with this is that if it cannot be determined, no information about the stability of the system will be available. Generally, typical Lyapunov function candidates are available for different model tasks, which can be "adapted" to the given task [4].

"Fixed-Point Iteration Methods" attempt to offer "alternative" control planning methods to circumvent the application of the Lyapunov function technique. The essence of the method is to transform the control task, namely, the calculation of the control signal to be given by the control system, into the iterative solution of a fixed-point problem so that one step of this iteration can be performed during a control cycle of a digital controller. Ensuring the convergence of the applied iteration is based on Stefan Banach’s fixed-point theorem. The given control task can be "transformed" into a fixed-point task in several ways, RFPT ("Robust Fixed-Point Transformation") offers a possible solution for this. Later, it became clear that the operation of this method, is related to the Lyapunov function-based technique [5].

The foundations of the method were developed in 2009 [6] [7]. RFPT is an iterative control method based on the fixed-point theorem of Stefan Banach proved in 1922 [1]. The procedure uses the available, usually inaccurate dynamic model of the system to be controlled to try to implement a trajectory tracking strategy based on purely kinetic/kinematic considerations by calculating the control forces. As the model used is inaccurate, the force calculated from it does not realize the desired motion. By observing the realized motion, the method "deforms" the input of the inaccurate model until the realized motion approaches the kinematically prescribed one well enough.

MRAC (Model Reference Adaptive Control) [15] [16] controllers are well suited for nonlinear systems. The model is based on tuning control signals instead of parameters. The essence of the technique is to compare the dynamic behavior of a feedback system to a reference model and control takes place accordingly. RFPT is also suitable for this [8] [9] without the use of Lyapunov functions. In the knowledge of the inverse of the model to be controlled, the input signals for the expected behavior can be determined as follows:

- r_d - expected behavior
- u - input signal

where $u = \phi(r_d)$. The inverse model is usually incomplete, in other words, it gives a different result than expected. The controller uses a deformation function to modify the input signal to compensate for this deviation. Based on [6] in case of SISO system:

$$G(r|r^d) \stackrel{\text{def}}{=} (r + K)[1 + B \tanh(A[f(r) - r^d])] - K \tag{1}$$

$$G(r_*^d|r^d) = r_*^d, \text{ if } f(r_*^d) = r^d \tag{2}$$

$$G(-K|r^d) = -K \text{ if } r_*^d = -K \tag{3}$$

where:

- $\mathrm{r}_*^{\mathrm{d}}$ is the deformed signal constituting the solution to the task
- K, A, B are the parameters of the controller

There is usually no specific suggestion for the choice of controller parameters, as the response function $f(r)$ also depends on the approximation model parameter used and the actual behavior of the system under control. Based on Banach's theorem, we should aim to try to give a contractive mapping of the function G in the proximity of the solution, i.e. close to r_*. To do this, the r-based derivative of G must be reduced to a value less than 1 in absolute value, which would require information about the derivative of the function $f(r)$.

The value of B can therefore be +1 or -1 according to the sign of the derivative of $f(r)$, the value of K should be chosen as a large number in comparison with which the values of r in the sum $(K + r)$ are small, and the value of A must be reduced until the iteration becomes convergent. (A very small A value causes slow convergence and inaccurate adaptability.)

In addition to the deforming function, a kinematic block forms part of the controller. Within the kinematic block, the difference between the realized trajectory and the prescribed path is determined as well as the derivative of the error and its integral as follows:

$$e_{int} = \int_{t_0}^{t} \big(q_n(\tau) - q(\tau)\big) \mathrm{d}\tau \tag{4}$$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} + \Lambda\right)^{n+1} e_{int} = 0 \tag{5}$$

where: Λ is a control parameter

The block diagram of the robust fixed-point transformation-based control is shown in the Figure 1:

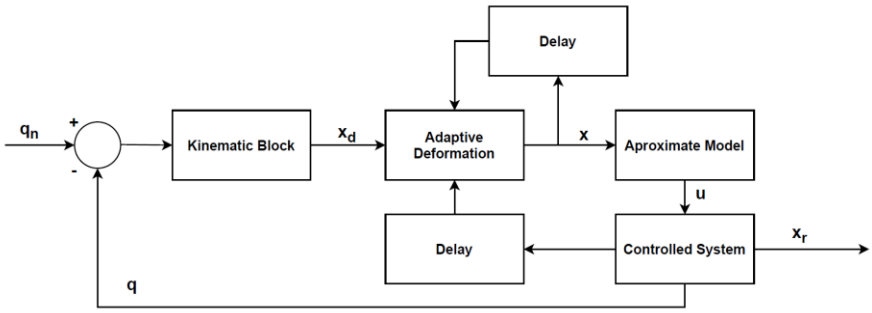


Figure 1
The block diagram of the robust fixed-point transformation-based control

The controller contains a total of four parameters:

- Λ, K, A, B

There is no specific method for determining and tuning these parameters. The value of the parameter $\Lambda \approx 1/\tau$ is roughly determined by the dynamics of the trajectory to be followed. A rapidly changing trajectory with very sluggish dynamics (τ is roughly a time constant) cannot be accurately tracked in the first place. If the dynamics of the track to be followed changes over time, it may also be necessary to tune this parameter.

Many tuning methods have been developed over several decades, for example, in order to set the PID controller correctly e.g. Ziegler-Nichols method (1942), which can be applied to processes where it is possible to operate the control cycle at the limit of stability, or the Oppelt method, which can be used to determine the individual parts based on the response of the process to step input. These methods have been developed for Linear Time-Invariant (LTI) dynamic models [13], mostly using the frequency image.

Adaptive RFPT presupposes strongly nonlinear system models in the first place, so in this case, thinking in frequency image and step inputs is not relevant (it is not advisable to conflict a nonlinear system with step inputs), so it does not have a tuning technique applicable to LTI models.

However, there are a number of optimal search methods that can also be used to tune controllers (e.g., PIDs), e.g. the genetic algorithm (GA) [16] [17] [18] [19]. Using this method, I developed a possible autotuning method for tuning adaptive RFPT.

2 Genetic Algorithm

The genetic algorithm (hereafter GA) is a rather widespread evolutionary strategy, the foundations of which were developed in 1975 by John Holland [14]. It has been used in many fields in recent decades: graph algorithm, extreme value problem, game theory, etc. However, it can also be used in case of tuning regulatory algorithms e.g. PIDs. Following a similar principle, I also developed a procedure for RFPT auto-tuning.

GA algorithms strive to ensure that the result of solving a problem is optimal within a given error threshold and accuracy. They are suitable for global optimization. They are sensitive to local optima, due to their stochastic nature, however, they can handle adequately this problem.

GA is an iterative process based on one initial population. Each population is made up of several individuals. These individuals are possible solutions to the problem under examination. However, there may be identical solutions among them. Individuals can be further broken down into genes. Genes represent the characteristics of an individual. In the process, repeatedly, new populations are created with each iteration. These are called generations. Each new generation is created from the current population using different selection, recombination, and mutation algorithms [10].

2.1 Individuals

In the case of GA, the individuals contain possible solutions to the given problem. There may even be redundancy between individuals. The first step in the process is to design the way individuals are represented. In many cases, an individual is described by a sequence of bits, a vector, or some structure. The parameters of the representation mode, the parameters giving the properties of the individual, are called genes. The algorithm modifies these genes during each iteration based on a given strategy.

2.2 Populations

The first iteration of the genetic algorithm is based on an initial population. The structure of the initial population depends on the particular problem. In many cases, it consists of a few hundred or a few thousand individuals. In the case of an optimum search problem where the location of the optimum can be guessed, this is taken into account when constructing the initial population. In this case, more individuals are created in the search area around the assumed location, otherwise, the individuals are evenly distributed. The reason for this is that by helping GA, the chances of finding a global optimum can be increased and the search time for the optimum can be reduced.

With each iteration, new populations are created from the populations. These are called generations. A generation is created using the current population, based on different selection, recombination, and mutation algorithms. Each individual is described by a "Fitness Function". During selection, aptitude is determined by this function.

2.3 Fitness Function

The individual goodness of each population is determined by a fitness function. The better this value, the closer the solution approximates the expected result, i.e. the global optimum. One of the most difficult tasks is to determine this function. It is one of the main building blocks of GA. It must be chosen with great care for each problem. During each iteration, when new generations are created, individuals with better fitness values are more likely to enter the new population (this may change depending on the selection procedure).

2.4 Selection

During selection, the individuals the genes of which we would like to be further inherited are selected from the current population based on their aptitude and quality. This is one of the conditions for the algorithm to converge towards the global optimum during the iterations. Selected individuals are called parents. There are several strategies for selecting the best individuals. These include random selection, fitness proportional selection, competitive, etc. The first two of these are presented.

2.5 Random Selection

It is less effective, however, it is one of the simplest procedures. In a given iteration, the parents are randomly selected from the given population with the same probability. Considering the Darwinian theory, this solution is not the most obvious, because in this case, the basic idea is that during the various developments, the weak individuals become extinct, and the strongest and most capable ones continue to reproduce. Regardless of fitness, all individuals are equally likely to be parents, thus further reproducing their genes, whether good or bad. The method can be further refined in order to be improved. With each iteration, it can be taken into account that if an individual has already been selected, it cannot participate as a parent later.

2.6 Fitness Proportional Selection (Roulette Method)

The method is based on the fitness function. A parent is selected from the current population in a way that the probability of selecting individuals in each population

is proportional to the fitness value. The same instance can be re-selected at each step. The chance of selection for a given individual is:

$$p(\hat{e}) = \frac{F\{\hat{e}\}}{\sum_{e \in p} F(e)} \tag{6}$$

where $F(e)$ means the fitness value of individual e. The roulette wheel method is usually used for selection. The higher the fitness value of a particular individual, the larger the slice you get from the wheel. Rotate the wheel to select a perimeter point. The higher the fitness value of an individual, the more likely it is to be selected by this method.

2.7 Recombination

After selection, the first step in creating new individuals is recombination. Upon recombination, a new individual is created from the selected parents. The new individual inherits the genes of the parents. This is called a crossover. There are several crossover methods, e.g. 1-point crossover, uniform crossover, intermediate recombination, heuristic crossover, etc.

2.8 Single-Point Crossover

In a single-point crossover, an individual is cut at a random point. The parent shall have n-genes and i will be a random number where $1 \pm i < n$. The new individual inherits its genes from one parent from 1 to i and from the other parent from $i + 1$ to n. In this case, the parents are divided into two parts, the individual parts being transferred one by one to the new individual. Continuing with the method, it is possible to cross the two parents not only at one but also at several points. From one parent, n genes are selected, which are transferred to the new individual, and then the missing genes are inherited from the other parent.

2.9 Smooth and Intermediate Crossover

Smooth and intermediate crossovers are based on the same logic. Each gene in the new individual will be one of the same genes in the parents. Each gene is selected with a 50% chance, i.e. half of the genes are exchanged between parents. In the case of an intermediate crossover, the value of the inherited gene changes, the value of which is described by a function. The latter method tries to achieve its larger variety. There is no general method for selecting individual crossover operators. In each case, it needs to adapt to the task. In most cases, each gene is independent. If there is a relationship between the genes, a special “intelligent operator” must be used.

2.10 Mutation

The search space is made possible by the mutation operator. However, this operator should be used with caution. If the genes of an individual are heavily modified, traits inherited from parents with good fitness values can be impaired. Depending on the data describing the individuals (bit sequence, vector, structure), there are several mutation operations (random element per gene or element permutation, inversion operator, neighborhood mutation, sequential mutation, etc.). It is very important that during the mutation, the operator should only change the value of the gene to such an extent that it does not leave the search space. For many generations, if the algorithm approaches the optimum, convergence to the optimum may fail, if the mutation rate is too high.

2.11 Pseudocode of GA Algorithm

Algorithm 1 Genetic algorithm

```
1: Setting strategic parameters
2: t := 0
3: Creating P_0 initial
4: Initial evaluation of P_0 individual based on the fitness function
5: while Not Exit condition(P_t) do
6:     P_t* := Selection(P_t)
7:     P'_t := Recombination(P_t*)
8:     P_{t+1} := Reinstatement(P'_t)
9:     t := t + 1
10:    Calculate the fitness function
11: end while
12: Choosing individual with best fitness values
```

As a first step, the parameters are set and produce the initial population. The number of individuals in the initial population depends on the solution of the problem. That can mean a number from a few hundred to a few thousand individuals. The individual individuals, as I mentioned earlier, can be bit sequences, vectors, or some structures. In the next step, each individual is evaluated based on the fitness function. The algorithm is an iteration process. The exit condition of the process varies. The condition is usually met when a predetermined “error threshold” or iteration (generation) is reached. It is necessary to limit the iteration steps because there may be a case where the GA gets stuck and does not converge towards the optimum. Selection, recombination, and mutation take place within the cycle core. If the exit condition is met, the process stops, and the individual with the best fitness value in the last population is selected.

3 A Demonstration of the Automatic Tuning Method for GA-based Adaptive RFPT Controller Parameters

There is no exact tuning method for selecting the parameters of the adaptive robust fixed-point transformation-based control algorithm. The task of the parameters of the kinematic block and the deformation function in the control circle detailed in Chapter 1 is known, however, the choice of the values of each parameter can be determined experimentally only in a way supported by observations. The expectation for a control task in most cases is to follow a prescribed trajectory with the smallest possible error. Tracking error can be minimized by the optimal selection of each parameter. For RFPT, this means setting four parameters (Λ, K, A, B). The process of the developed method has been extended by one step compared to the genetic algorithm:

Algorithm 2 Genetic algorithm extension with RFPT simulation

1: Setting strategic parameters
2: $t := 0$
3: Creating P_0 initial
4: **Running RFPT simulations using the initial population**
5: Initial evaluation of P_0 individual based on the fitness function
6: **while** Not Exit condition(P_t) **do**
7: $P_t* := Selection(P_t)$
8: $P'_t := Recombination(P_t*)$
9: $P_{t+1} := Reinstatement(P'_t)$
10: $t := t + 1$
11: **Running RFPT simulations using the new population**
12: Calculate the fitness function
13: **end while**
14: Choosing individual with best fitness values

The method is based on a simulation block. Within the simulation block, the RFPT control block visible in figure is realized. The block consists of the following main components:

- Kinematics block
- Deforming function
- Approximating function
- Exact model

During the simulation, the goal for the system is to follow the predefined trajectory as accurately as possible. Genetic algorithm-based tuning uses the result of this simulation block.

3.1 Chromosome, Representation

The genotype assigned to each solution is made up of four numbers, i.e., an individual has four genes. The alleles assigned to the genes correspond to some parameters of the RFPT (Λ, K, A, B). The phenotype generating algorithm is provided by none other than the RFPT simulation block using the genotype.

3.2 Setting Strategic Parameters

After chromosome representation, the first step is to determine the initial population. The parameters of the individuals in the initial population are determined in consideration of the search space. Each allele, as an allele within a different range, is given a value by generating evenly distributed random numbers. Each domain is as follows:

- Λ: $0.1 < x < 6$
- K: $1e2 < x < 1e13$
- A: $1 > x > 1e-6$
- B: $[-1,1]$

As the parameter space is too large for the initial population, except for Λ, the possible values can only be multiplies of 10. There is no specific method for determining population size [11]. Most of the time it depends on the problem. In this case, the simulation was performed with populations of 50 to 450 individuals. Each new generation is created from the elite of the current population. I selected a 10% elite rate [12].

3.3 Crossover and Mutation

The first step in the crossover is to choose the parents. The parents come from the elites. I chose single-point crossover as the crossover operator. Each parent has 4 genes. I used a randomly distributed random number generator to determine the point of intersection. $P_1 = [\Lambda_1, K_1, A_1, B_1]$ and $P_2 = [\Lambda_2, K_2, A_2, B_2]$, as well as C_p should be a random number generated in the range $[1, 2, 3, 4]$. If for instance $C_p = 2$, then $C = [\Lambda_1, K_1, A_2, B_2]$. The last strategic parameter is the choice of mutation rate. As in the case of the initial population formation, I limited the values of each parameter, the mutation rate helps to broaden the search space again. I used a random mutation per gene. 6% of all genes in the population as well as the value of each gene can be modified proportionately by generating a random number between $\pm 10\%$.

3.4 Stopping Criteria

The generic algorithm is an iterative method. At the end of each iteration, it should be examined whether it is worth continuing the optimal search or not. Usually, the stop condition consists of two parameters, an expected fitness value, and a maximum iteration number. Optimally, the first condition is met, in which case the individual with the best fitness value in the last population provides the best solution to the problem. However, due to the heuristic nature of the algorithm, this is not guaranteed at all. In the case of the autotuning method developed by me, it is difficult or impractical to set a fitness value as a stopping criterion. In the case of control, the goal is to minimize trajectory tracking error or even reduce it to zero. For the latter physical systems, it can be concluded that it will never be satisfied. Since in reality, the interfering signals are limited, e.g. a system displaced from rest would have to be displaced with an unrealistically large intervention signal and then slowed down to follow the prescribed trajectory immediately. As this is not met, there will always be a tracking error. I determined only the maximum iteration number for the exit condition of the genetic algorithm. To determine this maximum number of iterations, I performed several tests, which I will detail in a later chapter.

3.5 Fitness Definition

The aptitude of each specimen is described by the fitness function. The global optimum of this function is to be found. In the case of GA-based RFPT tuning, the fitness function is determined from the results of the simulation performed for each iteration. The goal is to minimize tracking error, ideally to zero. Achieving the latter is impossible, the reason for which has already been explained in the previous chapter. However, minimizing the error is a realistic goal.

It can be assumed that the smaller the error integral, the more accurately the system follows the required trajectory. However, as with most control algorithms (e.g., PID), the system can oscillate or even immediately diverge to infinity due to the selection of unsuitable parameters. As the genes of the specimens in the initial population are determined by evenly distributed random numbers, the latter case is also highly likely to occur. However, this does not mean that during further mutation, a specimen that initially provides an erroneous simulation result cannot converge to a good solution. In the articles below [23] [24], the authors used the following performance indicators to minimize the error generated during the simulations performed with each specimen:

- MSE: Mean of the Squared Error
- IAE: Integral of the Absolute Magnitude of the Error
- ITAE: Integral of Time multiplied by Absolute Error
- ISE: Integral of the Squared Error
- ITSE: Integral of Time multiplied by the Squared Error

$$MSE = \frac{1}{t}\int_0^t (e(t))^2 dt \quad (7)$$

$$IAE = \int_0^t |e(t)| dt \quad (8)$$

$$ITAE = \int_0^t |e(t)| dt \quad (9)$$

$$ISE = \frac{1}{t}\int_0^t e(t)^2 dt \quad (10)$$

$$ITSE = \frac{1}{t}\int_0^t t * e(t)^2 dt \quad (11)$$

Based on the performance indicators, the actual fitness value is determined based on the following equation [24] [25]:

$$Fitness\ value\ =\ 1/\ Performance\ index \quad (12)$$

The same performance indicators and fitness value calculations were used to tune the RFPT.

3.6 Dynamic Models used in Testing

I tested the tuning method using the following three nonlinear SISO type dynamic models:

- Van der Pol Oscillator
- Duffing Oscillator
- Inverted pendulum

The dynamic equation of the Van de Pol Oscillator is [20]:

$$m\ddot{q} - \mu(1 - q^2)\dot{q} + \omega_0^2 q + \alpha q^3 + \lambda q^5 = g \quad (13)$$

where, the values of each exact model parameter are:

- $m = 1.2$
- $\mu = 0.4$
- $\omega_0 = 0.2$
- $\alpha = 4$
- $\lambda = 0.3$

The dynamic equation of the Duffing Oscillator is [21]:

$$\ddot{x} + \sigma\dot{x} + \alpha x + \beta x^3 - \gamma\cos(\omega t) = u \quad (14)$$

, where the values of each exact model parameter are:

- $\alpha = 1$
- $\beta = 5$
- $\sigma = 0.02$

- $\gamma = 8$
- $\omega = 0.5$

The dynamic equation of the Inverted Pendulum [22]:

$$(m + M)\sin^2\theta\, l\dot{\theta} + ml\theta^2 \sin\theta\cos\theta - (m + M)g\sin\theta = -F\cos\theta \qquad (15)$$

Where the values of each exact model parameter are:

- $m = 0.8kg$
- $M = 1kg$
- $l = 0.6m$

The number of parameters of the approximating models (approx. model) within the adaptive RFPT control block was identical during each test, only the values of the parameters were modified compared to the exact model.

4 Experimental Results

The simulation was implemented in MATLAB software. I divided the simulation into an inner and an outer block. Inside the inner block is the RFPT block, which contains the kinematic block and the deformation function, as well as one of the exact and approximated models presented in Chapter 1. During each test, the controller must follow a sinusoidal trajectory. The genetic algorithm is built around the inner block. The genetic algorithm performs the following main steps for a given population number as shown in Chapter 3:

- Fitness calculation
- Selection
- Crossover
- Mutation

Only the iteration number was specified as the exit condition. To determine the optimal population size and the maximum number of iterations, I performed several measurements based on the following parameters:

- Population size: 50, 60, 70, ... 450
- Maximum iteration: 100

Additional parameters of GA:

- Elite rate: 0.1%
- Population mutation rate: 0.06%
- Gene mutation rate: 10%

For each population, I performed 100 tests, a total of 41×100 measurements. The measurements were performed on the same computer. In the following figures, I used the average of the measurements.

Figure 2 shows the integral of the error associated with the simulation of each population. The goal is to tune the RFPT so that during the simulation, the error integral is minimized (MSE, IAE, ITAE, ISE and ITSE produces similar results). The lower this number, the more accurately the controller works. It can be seen that after the 50th iteration no significant change occurs, for all population sizes RFPT-tuning is successful, the integral of the error approaches zero. As GA is heuristics-based, the larger the population, the more likely it is to reach the number of iterations sooner, after which tuning is no longer necessary. However, when choosing a population size, run time must also be taken into consideration, which increases in proportion to the size of the population.

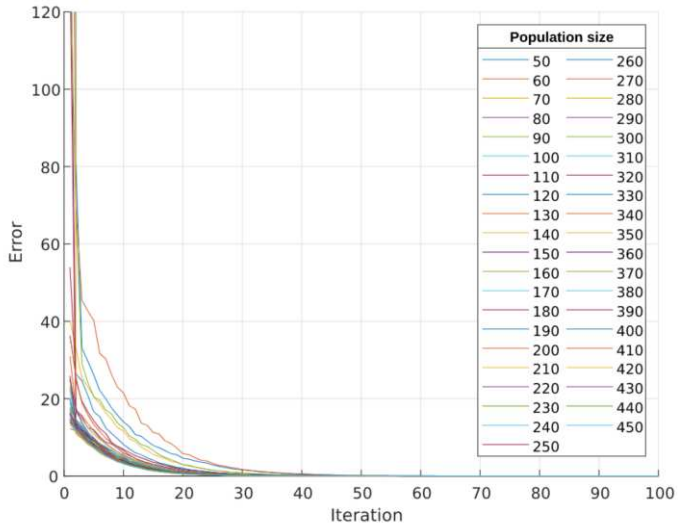
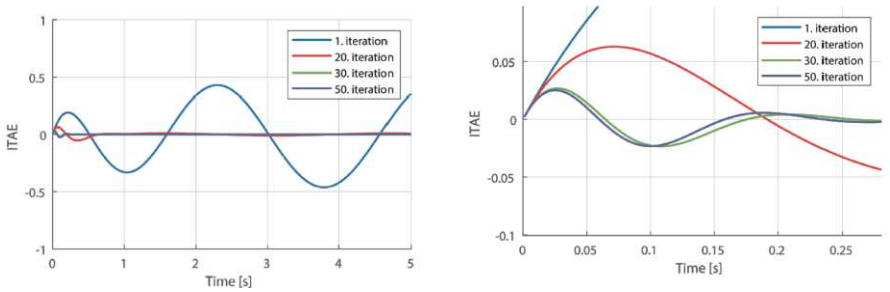


Figure 2
The integral of the error associated with the simulation of each population

Figure 3 (a) shows the process of tuning. In the case of the initial population, the system will most likely include specimens that may for instance lead to oscillations, but also, of course, those that already have the appropriate genes (RFPT parameters) at the beginning. Specimens with inappropriate parameters were not plotted, as the error rate was unrealistically large at the outset, so it was not used for plotting, but they were, of course, used for the performance of the genetic algorithm.



(a) ITAE simulation result

(b) Settling time and overshoot (details of Fig. (a))

Figure 3

Integral of Time multiplied by Absolute Error of the Van der Pol Oscillator

Figure 3 (b) shows that the trajectory tracking error decreases during tuning, at each time the controller is able to intervene faster, thus settling time and overshoot become smaller. However, the fact that in physical reality the intervening forces are limited should not be ignored. The further these values decrease, the greater the intervention signal required! This must be taken into consideration during tuning and a threshold value must be set for the magnitude of the interfering signal.

Figure 4 represents the result of the operation of the successfully tuned controller. The system had to follow a sinusoidal trajectory. The nominal trajectory is shown with a blue line while the tracking trajectory with a dashed red line. The nominal trajectory $q^N(t) = A_0 \sin(\omega_0 t)$, $A_0 = 3$ and initial state $q_0 = 0$, $\dot{q}_0 = 0$

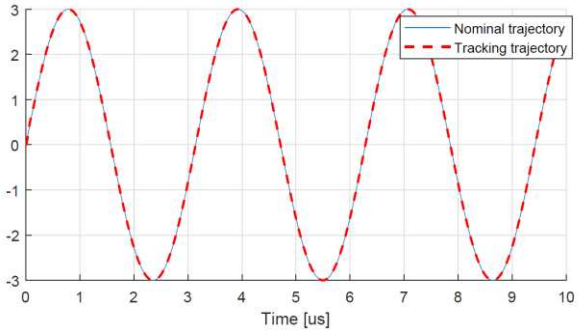


Figure 4

Trajectory tracking of the Van der Pol oscillator under tuned RFPT control

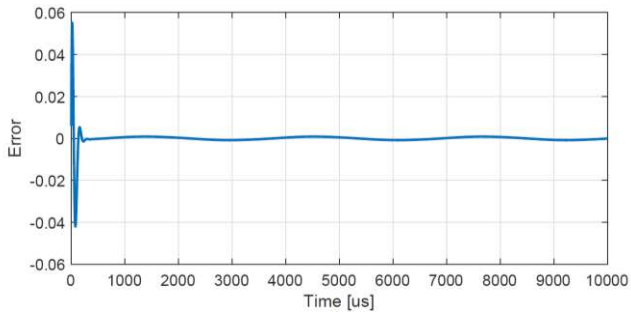


Figure 5
Trajectory tracking error of the Van der Pol oscillator under tuned RFPT control

Figure 6 shows the phase trajectory; Figure 7 shows the total control force while tracking of the Van der Pol oscillator under tuned RFPT control.

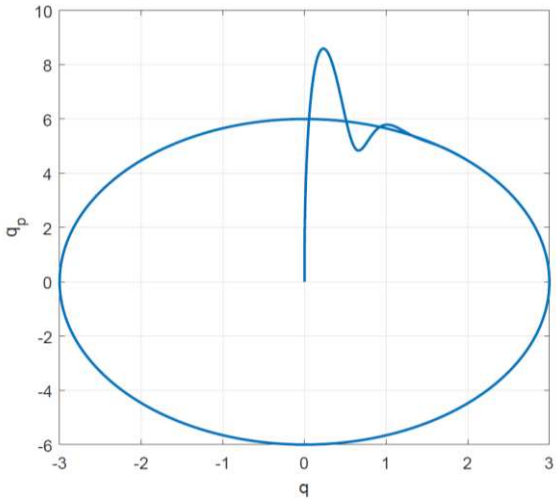


Figure 6
Phase trajectory tracking of the Van der Pol oscillator under tuned RFPT control

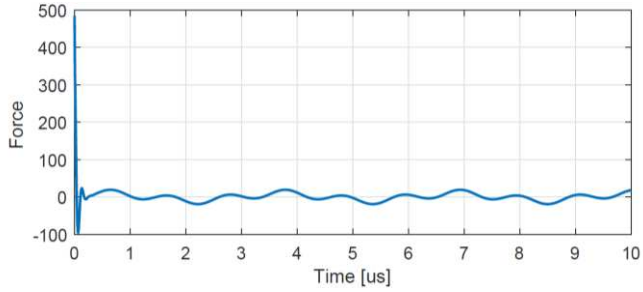


Figure 7
The total control force under tuned RFPT control

Conclusion

Based on the tests, it can be stated that the genetic algorithm is suitable for tuning the RFPT. Unlike PID, in the case of RFPT, there are no well-established, exact tuning methods, only individual parameters can be inferred from observations. Tuning can also be done manually, by monitoring the behavior of the system. However, the search area for each parameter is very large, it is possible that the right values can only be found after several attempts, however, this is not necessarily optimal. Thanks to GA, the whole operation can be automated, accelerated, and tuned correctly, under the right exit conditions. During the tests, I performed several measurements and tested the method on several dynamic models. In all cases, the controller was successfully tuned for the tracking error to nearly approximate zero.

References

[1] S. Banach: Surles opérations dansles ensembles abstraits et leur application aux équations intégrales (About the Operations in the Abstract Sets and Their Application to Integral Equations), Fund. Math., 1922, Vol. 3, pp. 133-181

[2] A. M. Lyapunov: A General Task about the Stability of Motion, Ph.D. Thesis, University of Kazan, Tatarstan (Russia), 1892

[3] A. M. Lyapunov: Stability of Motion, Academic Press, New-York and London, 1966

[4] P. Gahinet, P. Apkarian, M. Chilali: Affine parameter-dependent Lyapunov functions for real parametric uncertainty, In Proc. of Conference on Decision Control, 1994, pp. 2026-2031

[5] B. Csanádi, P. Galambos, J. K. Tar, Gy. Györök, A. Serester: Revisiting Lyapunov's Technique in the Fixed Point Transformation-based Adaptive Control, 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, 2018 June 21-23, doi: 10.1109/INES.2018.8523923

[6] J. K. Tar: Application of local deformations in adaptive control - a comparative survey (invited plenary lecture). In: Proc. of the 7th IEEE International Conference on Computational Cybernetics (ICCC 2009), Palma de Mallorca, Spain, November 26-29, 2009, pp. 25-38, 2009

[7] J. K. Tar, J. F. Bitó, L. Nádai, J. A. Tenreiro Machado: Robust Fixed Point Transformations in adaptive control using local basin of attraction. Acta Polytechnica Hungarica, 6 (1): 21-37, 2009

[8] J. K. Tar: Towards replacing Lyapunov's 'direct' method in adaptive control of nonlinear systems (invited plenary lecture). In: Proc. of the 2010 Mathematical Methods in Engineering International Symposium (MME 2010), October 21-24, 2010, Coimbra, Portugal, 2010

[9] J. K. Tar, K. Eredics: Simulation studies on various tuning methods for convergence stabilization in a novel approach of model reference adaptive control based on robust fixed point transformations. Acta Technica Jaurinensis, 4 (1): 37-57, 2010

[10] M. Mitchell: An introduction to genetic algorithms, Comput. Math. with Appl., Vol. 32, No. 6, p. 133, 1996

[11] Y. R. Tsoy: The influence of population size and search time limit on genetic algorithm, Sci. Technol. 2003. Proc. KORUS 2003, 7th KoreaRussia Int. Symp. Vol. 3, No. 1, pp. 181-187, 2003

[12] Sándor Szénási, Imre Felde: Configuring Genetic Algorithm to Solve the Inverse Heat Conduction Problem. Acta Polytechnica Hungarica, Vol. 14, No. 6, pp. 133-152, 2017

[13] Dániel Stojcsics, András Molnár: AirGuardian UAV Hardware and Software System for Small Size UAVs, Int J Adv Robotic Sy, 2012, Vol. 9, 174:2012

[14] J. Holland: Adaptation in Natural and Artificial Systems, The MIT Press [Reprint edition 1992] (originally published in 1975)

[15] P. Parks: Liapunov redesign of model reference adaptive control systems, in IEEE Transactions on Automatic Control, Vol. 11, No. 3, pp. 362-367, July 1966, doi: 10.1109/TAC.1966.1098361

[16] G. Kreisselmeier, B. Anderson: Robust model reference adaptive control, in IEEE Transactions on Automatic Control, Vol. 31, No. 2, pp. 127-133, February 1986, doi: 10.1109/TAC.1986.1104217

[17] X. Meng, B. Song: Fast Genetic Algorithms Used for PID Parameter Optimization, 2007 IEEE International Conference on Automation and Logistics, Jinan, 2007, pp. 2144-2148, doi: 10.1109/ICAL.2007.4338930

[18] X. Shao, L. Xiao, C. Han: Optimization of PID parameters based on genetic algorithm and interval algorithm, 2009 Chinese Control and Decision Conference, Guilin, 2009, pp. 741-745, doi: 10.1109/CCDC.2009.5191861

[19] G. Lin, G. Liu: Tuning PID controller using adaptive genetic algorithms, 2010 5th International Conference on Computer Science and Education, Hefei, 2010, pp. 519-523, doi: 10.1109/ICCSE.2010.5593559

[20] J. K. Tar, I. J. Rudas, J. F. Bitó, J. A. T. Machado, K. R. Kozłowski: Decoupled fixed point transformation based adaptive control of the generalized 2 DOF 6-type Van der Pol oscillator}, 2009 European Control Conference (ECC), Budapest, 2009, pp. 579-584, doi: 10.23919/ECC.2009.7074465

[21] J. Kabziński: Adaptive tracking control of a Duffing oscillator with hard error constraints, 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, 2016, pp. 1176-1181, doi: 10.1109/MMAR.2016.7575305

[22] T. A. Várkonyi, J. K. Tar, I. J. Rudas: Robust Fixed Point Transformations-based model reference adaptive control of inverted pendulums, 2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, 2011, pp. 591-596, doi: 10.1109/CINTI.2011.6108471

[23] Zwe-Lee Gaing: A Particle Swarm Optimization Approach for Optimum Design of PID Controller in AVR System, IEEE Transactions on Energy Conversion, Vol. 19, No. 2, pp. 384-391, 2004

[24] T. O. Mahony, C. J. Downing, K. Fatla: Genetic Algorithm for PID Parameter Optimization: Minimizing Error Criteria}, Process Control and Instrumentation 2000 26-28 July, University of Stracthclyde, pp. 148-153

[25] T. K. Teng, J. S. Shieh, C. S. Chen: Genetic algorithms applied in online autotuning PID parameters of a liquid-level control system}, Transaction of the Institute of Measurement and control 25, 5 (2003) pp. 433-450

Improved Image Denoising Using Wavelet Edge Detection Based on Otsu's Thresholding

Tuğba Özge Onur

Zonguldak Bülent Ecevit University, Dept. of Electrical-Electronics Engineering, 67100 Zonguldak, Turkey
E-mail: tozge.ozdinc@beun.edu.tr

Abstract: Images are affected by noise during their acquisition and transmission. Therefore, the denoising process is necessary to achieve higher quality images. However, both edges of the image and noise are characterized by high frequencies, loss of edge information may become unavoidable as a result of the denoising process. Thus, recovered, denoised images, become blurrier or less denoised. Therefore, a wavelet threshold denoising technique, based on edge detection, can be used to preserve more edge information and enhance the quality of the denoised image. In this paper, a novel image denoising method, based on wavelet thresholding by using Otsu's threshold, has been proposed and the clarity of the image which has been handled with this method is superior to that currently achieved by the other wavelet thresholds. The obtained results show that the proposed method, in this paper, provides better performance compared to commonly used wavelet image threshold denoising methods in terms of the visual quality of the denoised image. In addition, when the edge detection and denoising processes are combined, the deficiencies of the commonly used denoising methods are eliminated and a better denoising effect has been achieved.

Keywords: wavelet threshold; image denoising; wavelet edge detection; peak signal to noise ratio (PSNR)

1 Introduction

Image denoising is one of the basic tasks for the researchers dealing with image processing since there may occur distortions of images during the acquisition, processing, compression, transmission or reconstruction processes. Therefore, it is important to eliminate the noise from the images and increase the quality, or produce good estimates from noisy ones. The image noise can be Gauss, Poisson, or particle noise [1] [2]. The visuality and processing of the image are both affected by the noise. Therefore, it is aimed to preserve the useful information of the image and to reduce the noise by the image denoising process.

Since denoising is a preliminary process in the field of image processing, almost all researchers interested in image processing have dealt with this problem and

therefore researches on this effect made significant progress. Spectrum distribution is used for the traditional image denoising algorithms. In other respects, there have been various methods including the Markov field model, neural networks, non-linear diffusion, 3-dimensional block-matching, etc. to remove the noise from the images [3]. Although there are a wide variety of methods for the image denoising process, there have been challenges for most of the current denoising techniques because by the use of these techniques high frequencies can be suppressed. Thus, since noise can be characterized by high frequencies, high accurate results may not be achieved by the use of these techniques. Therefore, wavelet transform has been used extensively for various applications such as denoising in signal and image processing since there have been drawbacks of noise regularizing for other methods. In addition, successful results are achieved in applications due to an easily applicable algorithm and significant noise reduction effect of the wavelet transform. Due to its potential in the signal denoising process, its use has received significant attention from researchers. As more timely topics, the researchers have been attracted.

1.1 Related Works and Motivation

Although there have been various studies related to image denoising in the literature, researchers are still dealing with novel algorithms which have easy-to-implement applications. Lee et al. have presented a nonlinear diffusion filtering method and tried to improve the denoising process [4]. Mallat and Hwang [5] proposed an alternating projection method in 1992 and Zhu et al. [6] improved this method by obtaining the modulus maxima at each scale in 2017. In 2006, a denoising method based on sparse representation was used by Elad and Aharon [7]. Furthermore, artificial neural network models have been presented for the filtering processes of the noisy images [6] [7]. Moreover, fuzzy models can be used in noise image processing. Minh and Chen presented a generalized fuzzy system and achieved high performance for noise modeling in images [8]. A recurrent interval-valued fuzzy neural network was proposed by Juang et al. in 2011 [9]. Cheng and Juang proposed a fuzzy model that is based on support vector machine and margin selected gradient descent learning [10]. In addition, there have been several studies related to determining the efficient band by using a canonical correlation classifier. For example, Pozna and Precup suggested a new approach to model the system and represent the data with signatures [11].

Zall and Kangavari introduced an approach that was based on canonical correlation analysis and extracted correlation information between the significant paths [12]. Apart from these, denoising methods based on machine learning including k-nearest neighbor regression, etc. have been used for some researches [13]-[17]. Borlea et al. proposed a clustering algorithm that processes data sets of any size. They used it for the Iris dataset however suggested modifying it for different datasets [18].

Furthermore, not only due to the drawbacks of these methods but also the low performance of the classifier or model-based approaches in the literature, wavelet-based denoising methods can be available. Johnstone and Silverman [19] and Othman and Qian [20] have used wavelet transformation to distinguish useful information and noise from images. In addition, in recent studies, Bnou et al. have presented a new wavelet denoising method that uses an unsupervised learning model [21]. However, supervised learning models require prior learning of the corrupted image. Therefore, a thresholding filter is widely used to implement the denoising algorithm in the wavelet domain [22] [23]. Since the signal is transformed from the time domain to the time-scale domain in wavelet transform, a threshold value can be selected easily to reduce the noise. However, it is important to determine the appropriate value for the threshold, higher values result in better denoising whereas causes blurred edges. This is an undesirable result. Since the edges are mostly contain the information and basic character of the image, in some cases, loss or corruption of this information causes erroneous results [24]. Therefore, this results in data losses for the edges' knowledge since edges are characterized by high frequencies like noise, too. Donoho and Johnston [22] [24], [25] proposed hard and soft thresholding methods depending on the noise power and image size for denoising. In addition, other wavelet-based thresholding methods such as VisuShrink, Oracle Shrink, Normal Shrink have been used to obtain efficiency results in image denoising [26]. Thereafter, a wavelet edge detection based on both VisuShrink and scaled VisuShrink thresholding method was proposed by Liu and Ma and they proved that it outperforms classical wavelet thresholding methods [27]. Recently, machine learning is combined with traditional denoising methods [28]. However, the thresholding in denoising in images is still a crucial task and subject of research.

In recent researches, it has become even more important to reach the details of the images and eliminate the noise from them. In this paper, new wavelet thresholding to image denoising based on Otsu's thresholding is presented.

1.2 Contribution

The main contributions of this paper are as follows:

- The wavelet edge detection is used to detect the wavelet coefficients for the edges in the image before the denoising process by combining Otsu's thresholding which is one of the classical thresholding methods.
- Since the thresholds for the edges are set by Otsu's method, the wavelet coefficients can be thresholded without damaging the edges.
- The novelty of the proposed algorithm is the use of Otsu's thresholding and not to require any information concerning the noise level of the image. In addition, this is also the first study that combines Otsu's thresholding and wavelet transform algorithms to the best of the authors' knowledge.

- The extensive simulations are presented to validate the robustness and accuracy of the proposed denoising method. Based on the simulation results, it is revealed that the proposed method in this study is more effective compared with the other wavelet threshold denoising methods and it allows to achieve increasing denoising performances for the noisy images.

1.3 Organization

The rest of this paper is organized as follows. Section 2 describes the wavelet image edge detection. The proposed method in this paper is explained in detail in Section 3. Section 4 presents the simulation results to compare the performance of the method with related ones performed in recent years. Finally, the results are discussed and the paper is concluded in Section 5.

2 The Wavelet Image Edge Detection Method

In wavelet edge detection, a 2-dimensioned (2-D) wavelet transformation is required [9]. Therefore, two wavelets are needed as given in Eq. (1) to perform this transform.

$$\emptyset_{2^j}^x(x,y) = \frac{\partial \theta_{2^j}(x,y)}{\partial x} \quad , \qquad \emptyset_{2^j}^y(x,y) = \frac{\partial \theta_{2^j}(x,y)}{\partial y} \tag{1}$$

where $\emptyset_{2^j}^x(x,y)$ and $\emptyset_{2^j}^y(x,y)$ are the wavelets and $\theta_{2^j}(x,y)$ is a smoothing function. The wavelet coefficients of an image $g(x,y)$ can be determined as given in Eq. (2):

$$\begin{bmatrix} W_{2^j}^x g(x,y) \\ W_{2^j}^y g(x,y) \end{bmatrix} = \begin{bmatrix} g^* \emptyset_{2^j}^x(x,y) \\ g^* \emptyset_{2^j}^y(x,y) \end{bmatrix}$$

$$= 2^j \begin{bmatrix} \frac{\partial}{\partial x} g^* \theta_{2^j}(x,y) \\ \frac{\partial}{\partial y} g^* \theta_{2^j}(x,y) \end{bmatrix} = 2^j \nabla \left(g^* \theta_{2^j} \right)(x,y) \tag{2}$$

where 2^j is a scale factor. By considering Eq. (2), the edges can be detected with the scale factor 2^j since they can be explained by the local maxima of the gradient. If the image is distorted by additive noise, there may be some other pixels in the image which have a local maximum of the gradient. Therefore, it is important to be able to determine the coefficients belonging to the noise and the edges to prevent any loss of information in the denoising process. This distinction between the noise and the edges is made by considering Lipschitz exponent values if the additive noise in the image is the additive white Gaussian noise (AWGN) [29]. While applying

edge detection in the wavelet edge detection method, the image is an average filtered with appropriate length. For the second step, the initial is assumed to be $S_{2^0}g(x,y) = g(x,y)$ and wavelet transformation is applied to each row of $S_{2^j}g(x,y)$. Therefore, the used filters G_j and H_j in discrete wavelet transform (DWT) can be calculated as given in Eq. (3):

$$W^y_{2^{j+1}}g = \left(S_{2^j}g\right) * G_j$$
$$S_{2^{j+1}}g = \left(S_{2^j}g\right) * H_j \qquad \text{j=0, 1, 2, ...} \tag{3}$$

Here, the scale is represented by 2^j. For every row, coefficients of local maximum are found and saved in $W^y_{2^j}g(x,y)$. Since noise has low Lipschitz exponent values, the coefficients with these low values are removed in $W^y_{2^j}g(x,y)$. By applying this process, the remain coefficients in $W^y_{2^j}g(x,y)$ correspond to the edges of each row. If the same procedures are applied for each column, the coefficients of edges in $W^x_{2^j}g(x,y)$ can be obtained. Finally, a threshold value is determined by using these wavelet coefficients. However, Liu and Ma have proposed a wavelet edge detection method to overcome the problem of thresholding the wavelet coefficients with the appropriate threshold value. They used the wavelet edge detection method to detect the edges of the image before the denoising process. So, they set thresholds based on noise variances by using the VisuShrink threshold and since wavelet coefficients are protected, the information of the edges is not damaged [27]. However, determining the strength of noise in this thresholding may pose a problem for the images that have an indefinite noise level.

3 The Proposed Denoising Method: The Combination of Otsu's Thresholding and Wavelet Edge Detection

In the wavelet edge detection method, it is important to determine the appropriate threshold value while thresholding wavelet coefficients because noises are not clustered in a few wavelet coefficients. Therefore, if the threshold is not chosen high enough, the noise may not be reduced significantly. On the other hand, if it has a higher value, the better denoising performance will occur however it will result in blurred edges. So, to overcome this challenge, in the proposed method in this paper, the edges of the images are determined by the wavelet edge detection method in the first stage, and thereafter Otsu's thresholding is used to execute the wavelet threshold denoising process. The novelty of this study is to use Otsu's thresholding instead of thresholds which are based on the noise variances such as VisuShrink, Oracle Shrink, Normal Shrink thresholds, etc. The reason for using Otsu's thresholding is that Otsu's thresholding maximizes the between-class variance [30].

3.1 Otsu's Threshold Method

It is performed by selecting the lowest point between two classes. Therefore, edges can be determined without identifying the strength of the noise. Equations (4)-(8) describe the theoretical background of Otsu's thresholding.

$$\sigma^2 = \sigma_w^2(t) + q_1(t)[1 - q_1(t)][\mu_1(t) - \mu_2(t)]^2 \tag{4}$$

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \tag{5}$$

$$q_1(t) = \sum_{i=1}^{t} P(i)\ ,\ \ q_2(t) = \sum_{i=t+1}^{I} P(i) \tag{6}$$

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)}\ \ ,\ \ \mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{q_2(t)} \tag{7}$$

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)}\ \ ,\sigma_2^2(t) = \sum_{i=t+1}^{I} [i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)} \tag{8}$$

In Eqs. (4)-(8), $\sigma_w^2(t)$ is the weighted within-class variance, $q_1(t)$ and $q_2(t)$ are the probabilities of the classes, $\mu_1(t)$ and $\mu_2(t)$ are the means of the classes, $\sigma_1^2(t)$ and $\sigma_2^2(t)$ are the variances of individual classes [30].

3.2 The Proposed Image Denoising Approach

In the proposed image denoising method in this paper, firstly, the wavelet coefficients that correspond to the edges of the image are determined by using the wavelet edge detection method which is detailed in Section 2 with Eqs. (1)-(3). Thereafter, these coefficients are preserved and wavelet transform is performed to the distorted image by noise. As the third stage, the wavelet image threshold denoising process is applied by using Otsu's thresholding as defined in Eq. (9):

$$\widetilde{w} = \begin{cases} w & |w| \geq T \\ 0 & |w| < T \end{cases} \tag{9}$$

where T represents Otsu's threshold value. In the fourth stage, the coefficients which correspond to the image edges are replaced with the coefficients. In case the determined edges may also include noise, it is again thresholded with αT where α is an adjustment factor between 0.125 and 0.9. After the threshold T is determined, the α value is chosen to provide a better peak signal-to-noise ratio (PSNR). To determine the appropriate α value, some trials are performed by investigating the PSNR values and the results are shown in Table 1. The PSNR is defined as given in Eq. (10):

$$PSNR = 10 * \log\left(\frac{255^2}{\frac{1}{M * N}\left[\sum_{x=1}^{M}\sum_{j=1}^{M}\left(\left(g(x,y) - \hat{g}(x,y)\right)\right)^2\right]}\right) \quad (10)$$

where M*N denotes the image size, $g(x,y)$ and $\hat{g}(x,y)$ are the original and the reconstructed denoised images, respectively.

Table 1
PSNR (dB) values for different α adjustment factors and σ noise variances for 256*256 Chairs image

Noise variance	**PSNR(dB)**				
	α=0.125	**α=0.25**	**α=0.5**	**α=0.9**	**α=1**
σ =5	33.2851	33.4257	31.0618	28.4731	28.1641
σ =10	28.4524	28.5888	27.7786	26.5616	26.2353
σ =15	24.9799	25.2984	25.2109	24.5371	24.4315
σ =20	22.6096	22.9033	23.1415	22.7989	22.7463
σ =25	20.7398	20.9612	21.4797	21.3800	21.2821
σ =50	15.4561	11.4741	15.8323	16.4644	16.4117

It can be concluded from Table 1 that since noise is characterized by high frequencies like edges, in the case of the more disturbing image by noise α is chosen higher and otherwise smaller in the given range. Finally, the denoised image can be obtained by applying the inverse wavelet transform. These steps used in the proposed method are depicted with the flowchart in Fig. 1. In the proposed method given in Fig. 1, the wavelet coefficients which are the parameters of the model have been obtained by using Eqs. (1)-(3). Thereafter, the threshold value is determined with the Eqs. (4)-(8) and considering the adjustment factors and noise variances given in Table 1.

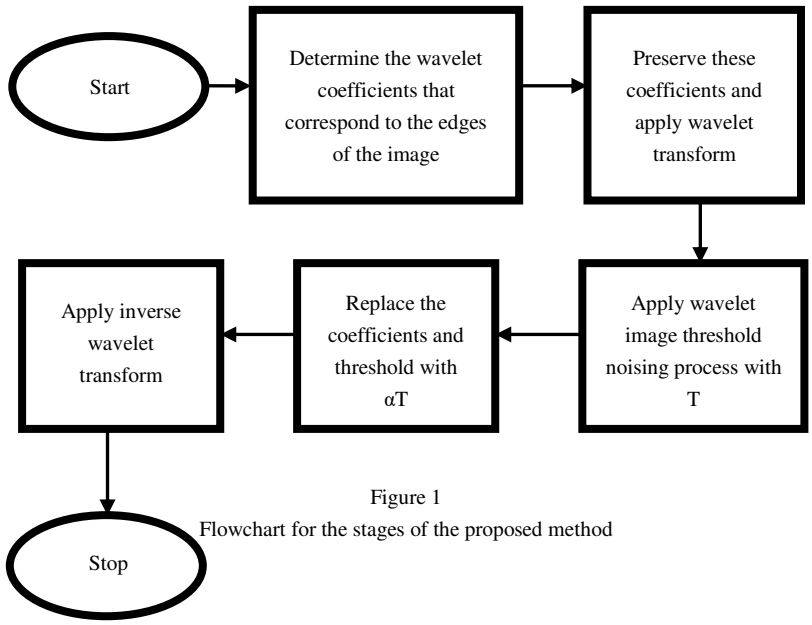


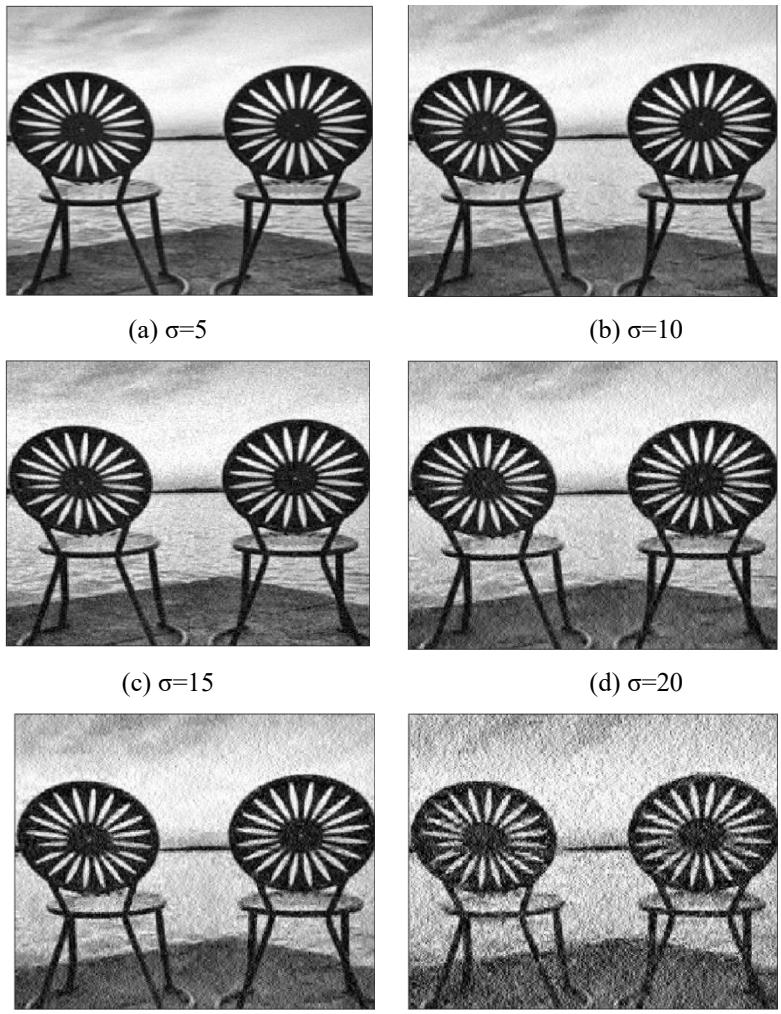
Figure 1
Flowchart for the stages of the proposed method

4 Subjective Evaluation Results

To evaluate the performance of the proposed method in this paper, Chairs image is preferred due to its different structured edges. This image is captured by a digital camera in Memorial Union which is a historical destination on UW-Madison's campus and resized in MATLAB 2018b program to provide convenience for the simulations. It is not easy to detect the different types of edges in the images completely by using a conventional edge detection method. In case of the image is corrupted by additive noise, this detection process will be even more difficult. For the simulations, the original 256*256 Chairs image is distorted by AWGN with zero mean and different variances. Figure 2 and Figure 3 show the original Chairs image and distorted ones, respectively. Simulations are carried out in MATLAB 2018b. To prove the validity and effectiveness of the presented method, the performance comparison is made with the method which uses both VisuShrink and scaled VisuShrink thresholds suggested by Liu and Ma in Ref. 21. Results shown in Table 2 compare the proposed method in this paper and the method in which thresholding is performed by both $T = \sigma\sqrt{2lnN}$ (VisuShrink threshold) and $T = \beta\sigma\sqrt{2lnN}$ (scaled VisuShrink threshold). The noise variances are determined to be in the range of 5-50 and the value for the α adjustment factor is determined as 0.25, 0.5, 0.9 for 5, 10, 15 and 20, 25 and 50 noise variances, respectively, by considering Table 1.

Figure 2
Original Chairs image

Figure 4 shows the obtained denoised images for different noise variances. In Fig. 4, figures in the first, second, third, and fourth rows belong to the denoised ones of the noise distorted with the noise variances 5, 10, 15, 20, 25, and 50 respectively. In addition, the left and right columns of each line correspond to the obtained denoised images of the proposed methods in this paper (Otsu's thresholding) and VisuShrink and scaled VisuShrink thresholding, respectively.



(a) σ=5 (b) σ=10

(c) σ=15 (d) σ=20

(e) σ=25 (f) σ=50

Figure 3
Distorted images by noise with different variances

It can be seen from Fig. 4 that even if the lowered VisuShrink threshold is used, the denoised images are getting notably blurred. When the proposed method with Otsu's threshold in this paper is used, smoother denoised images are obtained, and also more edge information is maintained.

(a)

(b)

(c)

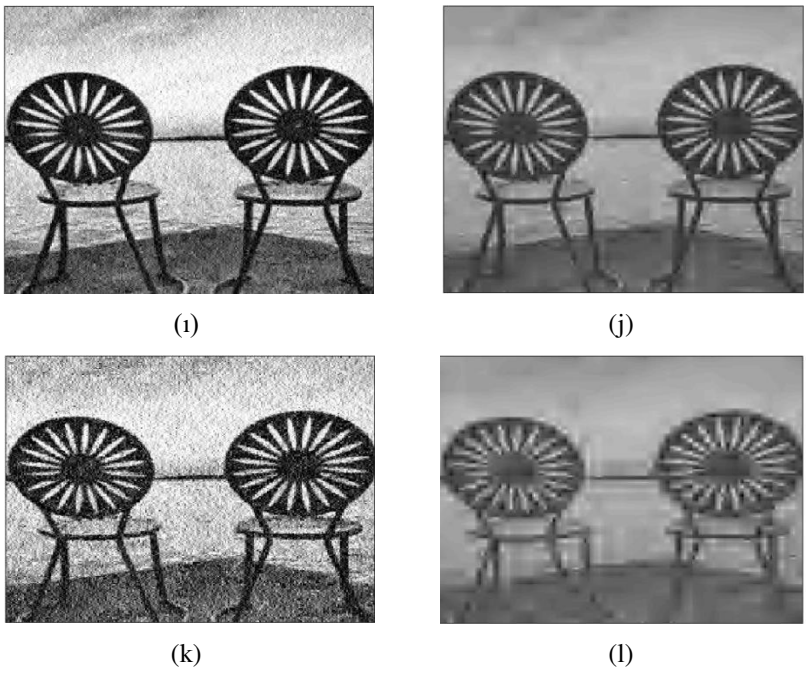
(d)

(e)

(f)

(g)

(h)



(ı) (j)

(k) (l)

Figure 4

The obtained denoised images: (a),(c),(e),(g),(ı),(k) are the results of the proposed method and (b),(d),(f),(h),(j),(l) are the results of the both VisuShrink and lowered VisuShrink thresholding with β=0.2

Table 2 presents the performance evaluation of the proposed method by comparing it with both VisuShrink and lowered VisuShrink thresholding according to the *PSNR* value which is calculated as given in Eq. (10). From Table 2, it can be seen that there is an increase for PSNR up to 0.0656~1.1949 dB especially with higher noise variances for the proposed method in this paper. In addition, the dilemma of the determination of the threshold value in the classical denoising methods is overcome by using Otsu's thresholding value. In other respects, the problem of the determination of the noise variance in the Liu and Ma method presented in Ref. 21 has also been eliminated by this proposed method. The theoretical analysis and simulation results obtained in this paper indicate that the proposed method can protect more useful information of the image and also provide more denoised images compared with the commonly-used wavelet threshold denoising methods. Therefore, it can be concluded that the proposed threshold method denoising effect is better than other wavelet threshold denoising ones.

As given in Table 2, when the same design specifications and noise variances are used for Chairs image, the performance of the thresholding in the proposed method provides higher PSNR values and hence more qualified and denoised images.

Table 2
Performance evaluation of the proposed method for Chairs image

Noise variance	PSNR(dB) for the proposed method	PSNR(dB) for Liu and Ma method
$\sigma = 5$	33.4257	33.3601
$\sigma = 10$	28.5888	27.9693
$\sigma = 15$	25.2984	24.9824
$\sigma = 20$	23.1415	22.8497
$\sigma = 25$	21.4797	20.2848
$\sigma = 50$	16.4644	15.3829

Conclusions

Image denoising is one of the most important applications in image processing. Using the knowledge that high frequencies characterize noise as well as edges, the denoising process and edge detection can be combined. Thus, deficiencies in commonly used denoising methods can be overcome. Although many denoising and edge detection methods are used today, different methods can be useful in different noise and image types. In this paper, a comprehensive framework for the image denoising method based on wavelet thresholding, by using Otsu's threshold, is provided. The appropriate adjustment factor for threshold value is determined by carrying out some trials for PSNR values and validation is done by applying the stages of the proposed method given in Fig. 1. Thereafter, a comparison between the proposed method and Liu and Ma [21] model, for the Chairs' image, is performed by using the same performance specifications. One can easily see from the obtained results that the denoising performance of the wavelet threshold denoising methods is effectively improved by the proposed method in this paper and it can be applied to different types of images. In the future additional research can be performed, relevant to this subject, by combining existing methods and then applied to different types of noise and images.

References

[1] G. Auber and P. Kolrnprobst: *Mathematical Problems in Images Processing. Partial Differential Equations and the Calculus of Variations* (2006) New York, NY: Springer Press

[2] A. K. Boyat and B. K. Joshi: "A review paper: noise models in digital image processing," *Signal Image Process. Int. J. (SIPIJ)*. 6(2) (2015) 63-75

[3] L. Fan et al.: "Brief review of image denoising techniques," *Vis. Comput. Ind. Biomed. Art.* 2(7) (2019) 1-12

[4] B. Lee et al.: "Harmonic decomposition in PDE based de-noising technique for magnetic resonance electrical impedance tomography," *Biomed. Eng.* 52 (2005) 1912-1920

[5] B. S. Mallat and W. L. Hwang: " Singularity detection and processing with wavelets," *Inform. Theory* 2 (1992) 617-643

[6] L. Zhu et al.: "Network-based method for mining novel HPV infection related genes using random walk with restart algorithm," *Mol. Basis Dis.* 6 (2017) 2376-2383

[7] M. Elad and M. Aharon: "Image De-noising via learned dictionaries and sparse representation," in Computer society Conference on Computer Vision and Pattern Recognition (New York, NY) (2006) 1-6

[8] N. M. Thanh and M. Chen: "Image denoising using adaptive Nneuro-fuzzy system," *IAENG International Journal of Applied Mathematics* 36(1) (2007) 1-7

[9] C. F. Juang, Y. Y. Lin and R. B. Huang: " Dynamic system modelling using a recurrent interval-valued fuzzy neural network and its hardware implementation," *Fuzzy Sets and Systems* 179(1) (2011) 83-99

[10] W. Y. Cheng and C. F. Juang: "A fuzzy model with online incremental SVM and Margin-selective gradient descent learning for classifications problems," *IEEE Transactions on Fuzzy Systems* 22(2) (2014) 324-337

[11] C. Pozna and R. E. Precup: " Applications of signatures to expert systems modeling," *Acta Polonica Hungarica* 11(2) (2014) 21-39

[12] R. Zall and M. R. Kangavari: " On the construction of multi-relational classifier based on canonical correlation analysis," *International Journal of Artifical Intelligence* 17(2) (2019) 23-43

[13] H. Noh et al.: "Deep neural networks by noise: its interpretation and optimization," *Machine Learn.* 11 (2011) 1-10

[14] S. Swami et al.: "Regularizing deep networks using efficient layerwise adversarial training," arxiv:1705.07819 (2017)

[15] K. Khera and A. Saini: " Image denoising using KD-tree and nearest neighbour based kernel regression model," IJISET 2(7) (2015) 410-414

[16] W. Li, D. Kong and J. Wu: "A novel hybrid model based on extreme learning machine, k-nearest neighbor regression and wavelet denoising applied to short-term electric load forecasting," *Energies* 10 (2017) 1-16

[17] D. Liu et al.: "When image denoising meets high-level vision tasks: a deep learning approach," in *Proc. of the IJCAI* (2018) 842-848

[18] I. D. Borlea et al.: "A unified form of fuzzy C-means and K-means algorithms and its partitional implementation," *Knowledge-Based Systems* 214(2-3) (2021) 106731

[19] I. Johnstone and B. Silverman: "Empirical byes selection of wavelet thresholds," *Annals Statist.* 4 (2005) 1700-1752

[20] H. Othman and S. E. Qian: "Noise reduction of hyper spectral imagery using hybrid spatial spectral derivative domain wavelet shrinkage," *Geosci. Remote Sensin.* 4 (2006) 397-408

[21] K. Bnou, S. Raghay and A. Hakim: " A wavelet denoising approach based onunsupervised learning model," *EURASIP J. Adv. Sig. Pr.* 36 (2020) 1-26

[22] D. L. Donoho and I. M. Johnstone: "Adapting to unknown wavelet shrinkage," *J. Amer. Statist. Assoc*. 90(432) (1995) 1200-1224

[23] P. Hedaoo and S. S. Godbole: "Wavelet based thresholding approach for image denoising," *IJNSA* 3(4) (2011) 16-21

[24] D. L. Donoho and I. M. Johnstone: "Ideal spatial adaptation by wavelet shrinkage," *Biometrika* 81(3) (1994) 425-455

[25] D. L. Donoho, "Denoising by soft-thresholding," *IEEE Trans. Inf. Theory* 41(3) (1995) 613-627

[26] F. Xiao and Y. Zhang: "A comparative study on thresholding methods in wavelet-based image denoising," Procedia Engineering 15 (2011) 3998-4003

[27] W. Liu and Z. Ma: "Wavelet image threshold denoising based on edge detection," in *Proc. CESA* (2006) 72-78

[28] M. Arora et al: "Wavelet denoising: comparative analysis and optimization using machine learning," in *Proc. 2014 9th International Conference on Industrial and Information Systems (ICIIS)* (2014) 1-6

[29] S. Mallat and S. Zhong: "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.* 14(7) (1992) 710-733

[30] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern. Syst*. 9 (1979) 62-6

Considering the Functioning of an e-learning System, Based on a Model for Assessing the Performance and Reliability of the System

Evgeny Fedkin, Saule Kumargazhanova, Saule Smailova, Natalya Denissova

D. Serikbayev East Kazakhstan Technical University
Faculty of Information Technology and Intelligent Systems
A. K. Protazanov Str. 69, 070004 Ust-Kamenogorsk, Kazakhstan
e-mail: {EFedkin, Saule.Kumargazhanova, SSmailova, NDenissova}@ektu.kz

György Györök

Óbuda University, Alba Regia Technical Faculty Budai út 45, H-8000
Székesfehérvár, Hungary, e-mail: gyorok.gyorgy@amk.uni-obuda.hu

Abstract: This work discusses the LMS (Learning Management System), as part of the University's Educational Portal and identifies the main hardware/software components of such an e-learning system. The conceptual model is presented, in the form of a queuing network. Analysis of the performance and reliability of the system, is used, as criteria. The Authors assessed the productivity of the EKTU e-learning system, based on the proposed model.

Keywords: learning management system; information and communication technologies; on-Line education; distance learning; educational environment; educational portal

Introduction

In a pandemic context, a global unplanned and one-time transition to distance learning technologies has become a necessity. The problems of heterogeneity of information systems, the lack of universal solutions and intersystem interaction, including learning management systems (LMS), have become aggravated.

A variety of LMS have been developed in recent years. These platforms often provide similar functionality, so choosing a learning platform is not an easy task.

There are many methodologies for assessing the quality of e-learning in the scientific literature. In works [1-4], are given the approaches to software development based on user preferences and experience, the "user-centered" design. In the last few years, systems with adaptive learning management have appeared. This is a new type of e-learning system based on automatic recognition and prediction of user preferences and self-adaptation to user requirements. In works [5-7], studies are presented based on social networks and communication between learners (students) and educators (teachers).

There are also studies [8] in the field of e-learning systems devoted to the pedagogical and economic aspects of systems, such as the formation of a knowledge system, the formation of a system of professional skills and abilities, the profitability of the educational process.

E-learning systems are being developed both as separate systems and as a part of educational process management information systems in educational institutions. Many universities implement systems that integrate different systems and subsystems. Since the main activity for the university is educational activity, the educational portal acts as such a unified system. Portals form a single entry point for various categories of users, including access to the e-learning system.

1 Designing of an e-learning System (LMS)

Modern LMSs are designed based on web technologies. The consumer properties of systems built on the basis of web technologies are determined by such characteristics as [4]: content, design, performance level, functionality, productivity, security level.

For system users, the main characteristics are performance (system response time to user requests) and system reliability (no failures when working with the system).

The performance and reliability of the system depend on the quantitative and qualitative characteristics of the server equipment of the information system:

- Number and speed of processors (CPU)
- Amount of RAM
- Bandwidth and the amount of disk space
- Bandwidth of network equipment

The most common performance metrics for web-based systems include:

- Response time when transferring data
- Response time of the system's website
- Bandwidth (requests or bits per unit of time)

- Number of errors per unit of time
- Number of visitors per unit of time

When considering the LMS as a system, [4] is recommended to classify certain service components rather than the services themselves. In general, the system consists of a set of databases, information processing facilities, information provision facilities and system interaction at the network level.

For a more convenient consideration of the components of the system, they can be divided into two large groups [4]:

- Application-level components that provide work with system data, implementing logical integrity and visual presentation;
- Base-level components that provide basic functions for application-level components: data transmission over the network, data security, data storage systems, etc.

In general, the e-learning system model can be represented as follows [8]:

{S, SW, HW, A, C, PN} (1)

Where, S - system web service infrastructure, SW - system software, HW - system hardware, A - system applications, C - system content, PN - communication networks.

Conceptually, the e-learning system model can be represented as a queuing network. To design the LMS architecture, the model needs to display a finite number of web servers, application servers, and database servers (Figure 1).

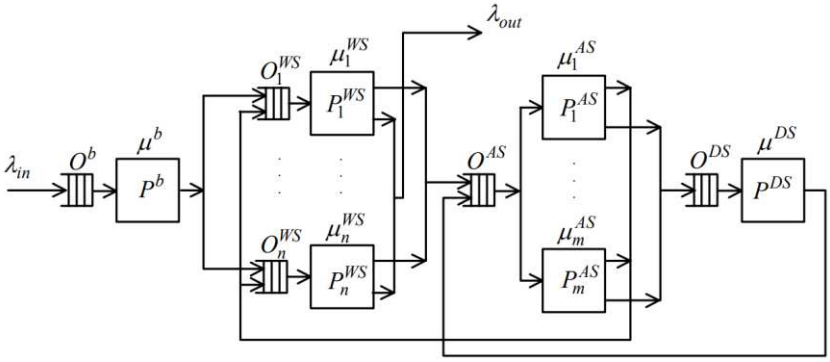


Figure 1
LMS conceptual model

The LMS work according to the given model is as follows:

- Input requests to the system arrive with intensity λ_{in} into the O^b queue of the network switch P^b, which dispatches the request.

The switch matches the virtual IP address with a valid web server address in a random time t_p with an intensity of $\mu^b = \frac{1}{t_b}$.

- The web server $P_i^{WS}, i = 1, n$ receives the request and generates an HTML page that is transmitted to the output stream - λ_{out} or sends a request to the application server $P_i^{AS}, i = 1, m$.
- The application server implements the business logic of the system and makes requests to the database server. The application server spends time $t_j^{AS}, j = 1, m$ with the intensity $\mu_j^{AS}, j = 1, m$, to execute the request. The functionality of this system component supports the system logic and is implemented in the form of CGI scripts or server applications. The result of the application server is the result of processing the request, which is sent to the web-server or a request to the database on the P^{DS} database server.
- The database server provides a system for storing system data and provides mechanisms for receiving data and updating them through a query language (SQL, stored procedures, etc.) based on transactions. A transaction to a database management system is implemented in a random time t^{DS} with an intensity $\mu^{DS} = \frac{1}{t^{DS}}$. The results of database queries are sent to the application server for further processing.

1.1 Assessment of System Productivity Based on the Model

Various indicators can be used to assess an e-learning system's productivity based on a web solution, which includes the average request processing time, average request size in bytes, average response size in bytes, rendering time of an HTML page in a browser, and others.

Since the central element of the e-learning system based on a web solution is the server part (web servers, application servers and database servers), the system productivity can be considered as the productivity of this particular central part of the system as a whole.

The following main indicators can be distinguished to assess the productivity of the system [9]:

- Intensity of requests (2), which characterizes the average number of requests that are simultaneously processed in the system. This indicator is determined by the number of users of the system and the number of requests that the user generates on average when working with the system.

$$\rho = \lambda \cdot t_{rst} \tag{2}$$

where λ is the intensity of the flow of requests; t_{rst} - request service time

- Probability of rejection (3) - the proportion of requests that are rejected. This indicator for a quality system should tend to zero. Modern systems based on web solutions have dozens and even hundreds of thousands of service channels, which means that the probability of failure for such systems is practically zero.

$$\rho_{rjc} = \frac{\rho^n}{n!} \cdot \rho_o \tag{3}$$

where n is the number of service channels, ρ is the intensity of requests, $\rho_o = \frac{1}{\sum_{k=0}^{n} \frac{\rho^k}{k!}}$ is the probability that the channel is free.

Nominal performance of the system (4) - the number of requests that the system can process per unit of time.

$$N = \frac{n}{\rho} \tag{4}$$

where n is the number of service channels, ρ is the intensity of requests.

The actual system performance (5) is % of system utilization of the nominal system performance:

$$F = \frac{\lambda}{N} \tag{5}$$

where λ is the intensity of the flow of requests, N is the nominal performance of the system.

The required size of the data transmission channel (6) shows the required speed of the data transmission channel through the network required for the system to work in the network.

$$S = \frac{\lambda \cdot r}{125 \cdot 60} Mbps$$

Where, λ is the intensity of the flow of requests, r is the average size of a response to a request in kilobytes, 125 kilobytes per 1 Mbps, 60 seconds per minute.

The presented model is the basis for developing e-learning systems and assessing their performance.

2 Implementation of an e-learning System Based on EKTU

As mentioned above, the main business processes at universities are those related to learning. To automate these processes, universities use heterogeneous systems that provide one or another aspect of work in this area. Portal solutions are used to ensure the connectivity of these heterogeneous systems in higher educational institutions.

In EKTU named after D. Serikbayev, since 2003, a similar solution has been developed [10, 11]. The educational portal is the central part of the university information system, which is shown in Figure 2.

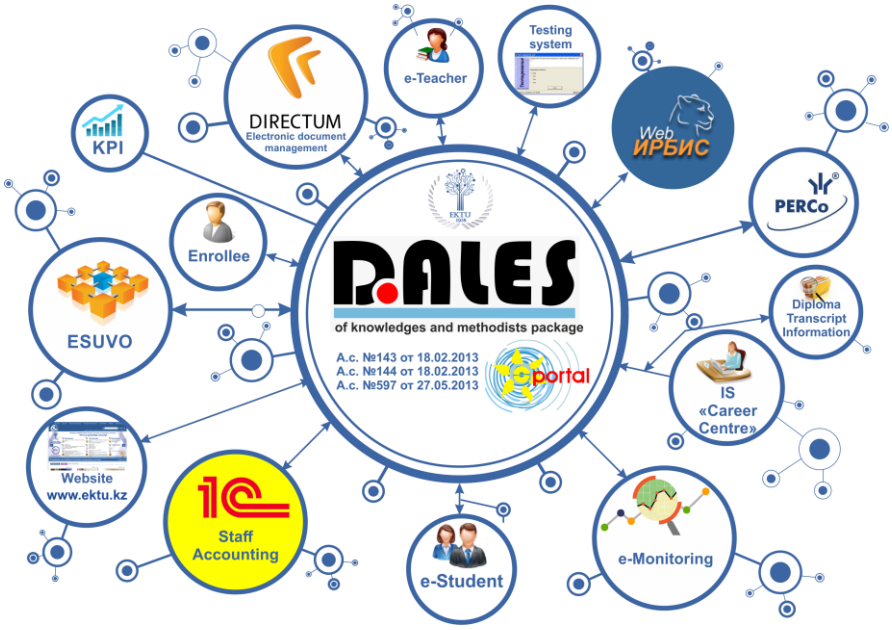


Figure 2

Scheme of the information system of EKTU

As can be seen from the scheme, the EKTU educational portal, being the central part of the information system, interacts with other information systems of the university - the IRBIS library system, access control system, accounting (1C), electronic document management system (Directum), the official website of the University and other systems.

2.1 Educational Portal of EKTU Named after D. Serikbayev

The architecture of the EKTU educational portal is built using client-server technology and is shown in Figure 3. The database management system Microsoft SQL Server 2017 is used to store data. Information is accessed using the website (http://www.do.ektu.kz) running under the control of a web server, the Internet Information Server 10.0, based on the Windows 2019 Server operating system.

The use of a thin client architecture, in which all components are located on the server, allows minimizing traffic, both from the client-side and from the server-side.

This scheme allows you to separate the command-control functions and the functions of providing and forming outputs. Due to the flexibility of the proposed scheme, you can make changes to one component without correction or with minimal correction of the other. In addition, the synthesis of the educational process control system is simplified since each block can be designed relatively independently, observing only the specifications of the interface between the blocks.

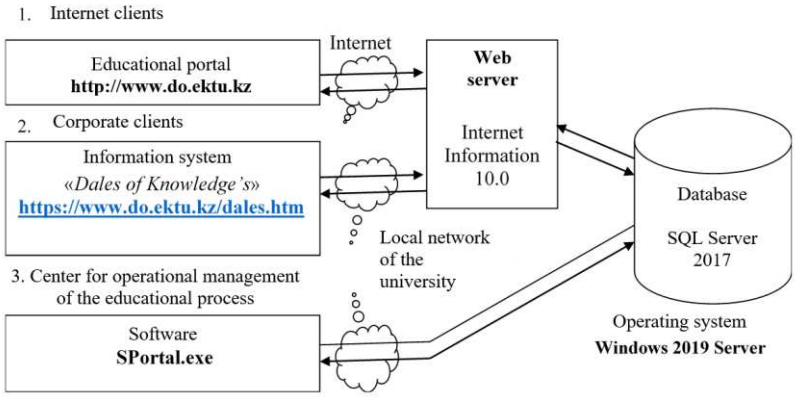


Figure 3
Educational portal architecture

Using the web interface to access the educational portal of the university provides many advantages in the form of versatility, the ability to work remotely, and interactivity.

Figure 4 shows the hardware and software structure of the EKTU Educational portal.

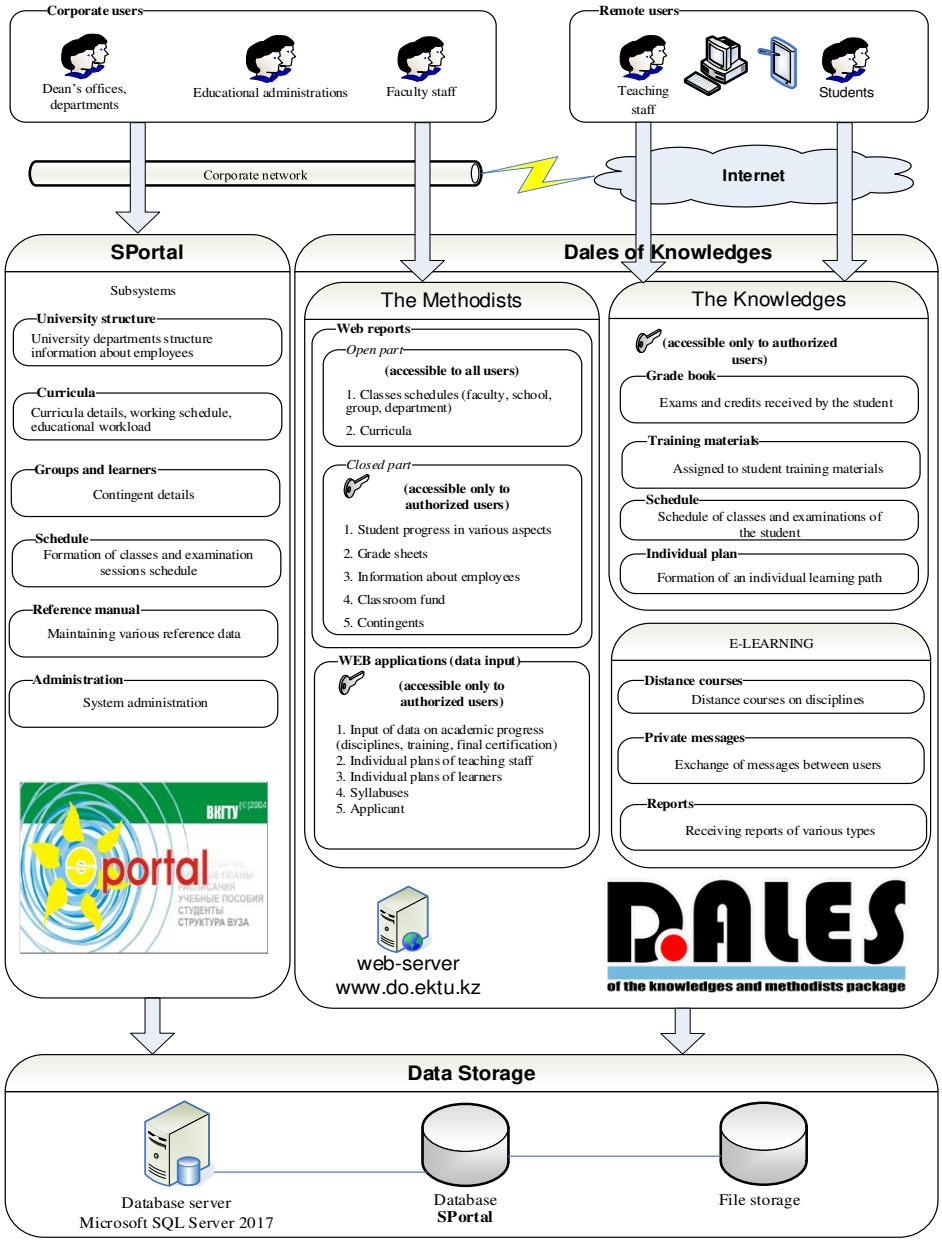


Figure 4
The structure of the EKTU portal

This structure includes the following components:

1. Database management system. This component is intended to ensure the functioning of a relational database for an educational portal. In our case, Microsoft SQL Server 2017 acts as a database management server. The main databases on which the university educational portal operates are

   a) SPortal is a central database that stores data related to the provision of the educational process (contingent, curricula, disciplines, teaching staff, etc.)

   b) File storage - a database designed to store binary data.

2. SPortal software package (fig. 4). This software package is a stand-alone windows-application for working with the database of the educational portal in the corporate environment of the university. SPortal is intended for the administrative services of the university - dean's offices, departments, educational management to maintain the data required for the organization of the educational process and various reference data. This software package includes such subsystems as:

   A) The structure of the university - designed to store the organizational structure of the university and the staff. This system is closely integrated with the "1C: Personnel" system from which information about the structural divisions of the university and their personnel is received.

   B) Curricula - designed to store information about various curricula on which training is carried out at the university, the formation and distribution of academic streams and teaching staff workload.

   C) Groups and students - designed to keep records of the university contingent

   D) Schedule - designed for scheduling classes and exams

   E) Reference books - designed to maintain various reference books that are required for the operation of various subsystems of the educational portal.

   F) Administration - designed to configure various subsystems of the educational portal.

This software package was developed using the Embarcadero Delphi 2009 programming environment. This environment has advanced tools for working with databases and a wide range of tools for displaying and editing data.

Interaction with the Microsoft SQL Server 2017 database server is based on ActiveX Data Object (ADO) technology using the Microsoft OLEDB Provider for SQL Server. The scheme of work of the application with this database is shown in Figure 5.

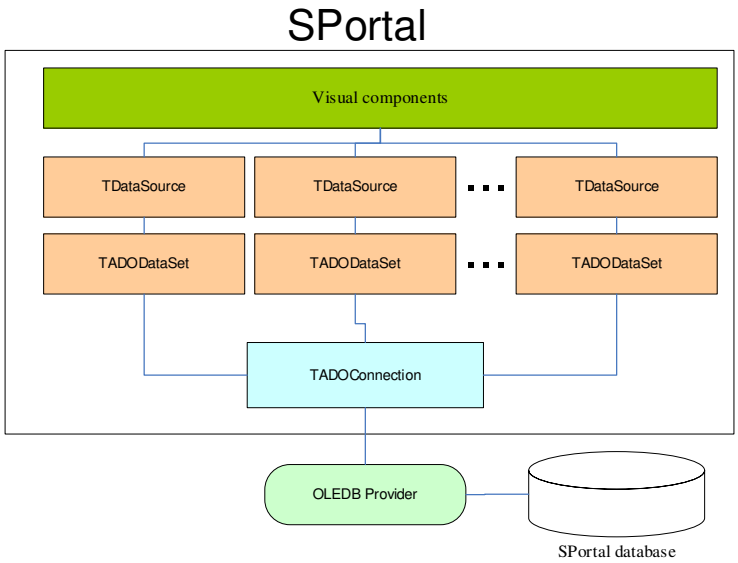


Figure 5
Scheme of work of SPortal software with a database

As you can see from the scheme shown in Figure 5, the TADOConnection component is used to connect to the database, which implements the connection to the database through the OLEDB provider. This component interacts with the TADODataSet components, which is a set of data in a database. The TDataSource components are used to interact with the data set in the TADODataSet and visual components. This component is responsible for retrieving data from the data source, passing it to the visual component for display and transferring updated data back to the dataset.

3. Dales: The Knowledges Web Services. These services are a set of web applications that provide an access point to various functionalities of the educational portal. These services are divided into the following categories:

   A) The Methodist - web services are designed for teaching staff and administrative staff to work with the educational portal. This part is divided into:

      - An open part that is accessible to all users without authorization. This part includes reports on class schedule and curriculum
      - A closed part, which is accessible only to authorized employees of the university. This part includes reports on progress, contingents, classroom fund, etc.
      - Web applications that are designed to enter data into the database of the educational portal by various employees of the university. This part includes journals of progress and attendance, grade sheets, data on applicants, individual plans of teaching staff and students, etc.

B) The knowledge - is a student's personal account, where he can get information about his progress, schedule, training plan, etc.

C) E-learning - is an e-learning system (LMS), which provides training on various courses for teaching staff and students.

Web services operate on a “request-response” basis (Figure 6).

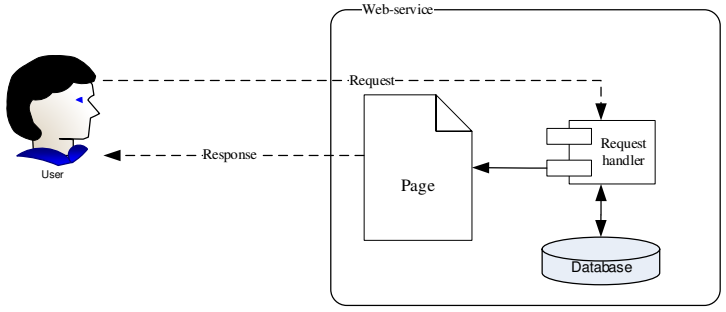


Figure 6
Scheme of the Web service operation

These web services are operated by the Microsoft Internet Information Services 10.0 web server. As technologies for implementation, technologies such as ASP (for the implementation of reports and data entry systems) and ASP.Net (for complex web applications, such as an e-learning system) are used.

Let us consider how the operation of the reporting system and the data entry system of web services based on ASP technology is implemented, which are implemented on the educational portal of EKTU. The implemented scheme of work based on this technology is shown in Figure 7.

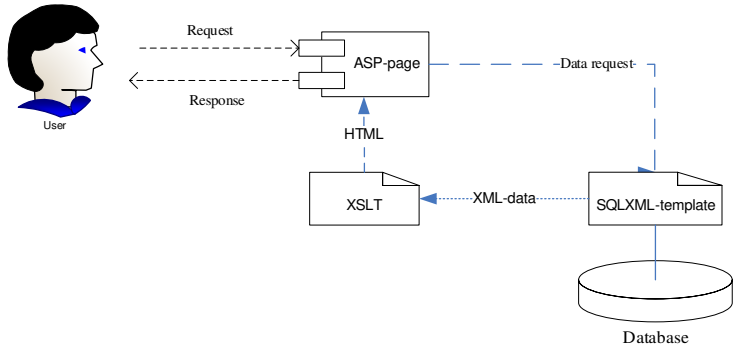


Figure 7
Scheme of the ASP-page of the educational portal operation

4. Directory services. This architecture component is intended for storing data about users, computers, user groups, group membership, and authentication

and authorization of system users. In our case, the directory service is the Microsoft Active Directory. This service provides data based on the ektu.kz domain.

2.2 E-learning System (LMS) as Part of the EKTU Educational Portal

The EKTU e-learning system is implemented on a single basis of the educational portal. From this portal, the system receives information about the contingent of students, information about active groups, data about existing teachers, data on disciplines, information on individual plans of students, data on current curricula. Based on the received data in the system:

- Thee e-courses are formed in groups, assigned to teachers, by semesters
- Various types of reports are created
- Export of progress to the educational portal is performed

Let us consider the software architecture of the server part of the distance learning system since it is it that determines the main functionality of the system. The software structure of the server part of the distance learning system is shown in Figure 8.

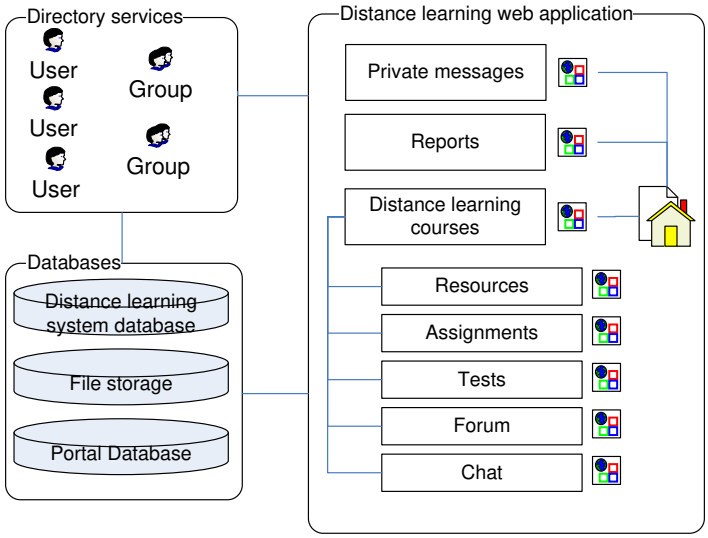


Figure 8
LMS software architecture

Consider the elements of the system software architecture presented in Figure 8.

- Directory services. This part is a repository of data about users, security groups and users' membership in them. This element of the architecture provides authentication and authorization of users of the educational portal and the distance learning system as part of the educational portal. Other elements of the system have access to this element, which makes it possible to provide a single point of safety.
- Database. In this part, we have three databases that support the functioning of the distance learning system. The databases operate on the basis of the Microsoft SQL Server 2017 database management system. The system has the following databases:

  a) Database of the educational portal. This database contains information from the educational portal of the university and contains such information as a catalog of disciplines, a catalog of curricula, student contingent, information about the teaching staff and other data.

  b) File storage. This database contains binary data representing files of various formats (Word, Excel, PDF, etc.). The file store is displayed in a separate store so as not to mix binary and relational data in the system.

  c) Database of the distance learning system. This database contains data that is intended for the functioning of distance learning. This database is linked to the database of the educational portal, which makes it possible not to duplicate data between the database of the educational portal and the database of distance learning.
- Web-based e-learning application. This part is a web application that provides the formation of dynamic HTML-pages based on asp.net 4.0 technology for displaying to users. In this part, we can distinguish the following groups of dynamic pages:

  a) Home page - this page is the entrance page of the distance learning system. This page displays ads posted in the system for various groups of users and contains transitions to other elements.

  b) Module "Private messages" - this module is designed to exchange messages between users. Messages in this module are in the mode of dialogue between users according to the 1: 1 scheme.

  c) "Reports" module - this module is a set of dynamic reports on the functioning of the distance learning system: data on courses, user activity, progress (успеваемость), unchecked assignments and other types of reports.

  d) Module "Distant Learning Courses" - this module is the main part of the distance learning system that provides the page operation of individual distance courses. This module contains the following submodules: resources, assignments, tests, chat.

2.3 Assessment of the Effectiveness of the EKTU e-learning System

Various indicators can be used to assess the efficiency and reliability of systems. For our model of an e-learning system based on a web solution, the main indicators for assessing the efficiency and reliability of the system are related to the number of users who work in the system at some point in time and the network traffic that is generated during operation. Consequently, we will assess the effectiveness of the VKTU e-learning system based on these indicators.

The EKTU e-learning system has the following hardware:

1) Web server: HPE ProLiant DL325 Gen10, AMD EPYC 7351P 16-Core Processor, 2.4 GHz, 96 Gb RAM, 6 * 1.5 Tb HDD, RAID 1
2) Database Server HPE ProLiant DL385 Gen10, 2 * AMD EPYC 7251 8-Core Processor, 2.1 GHz, 64 Gb RAM, 8 * 1 Tb HDD, RAID 1
3) Corporate network - 100 Mb
4) Network inside the server room - 1 Gb
5) Internet connections - 500 Mb

This hardware configuration allows serving a significant number of system users. The graphs (Fig. 10, 11) show the activity of users in the e-learning system (students and teaching staff).

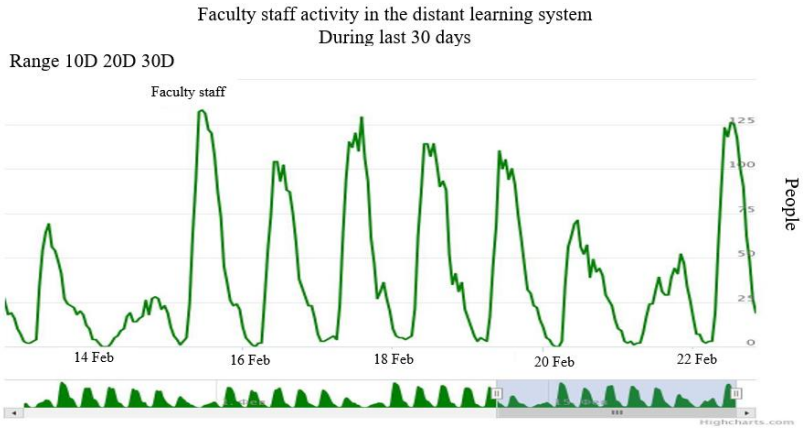


Figure 10
Faculty staff activity in the system

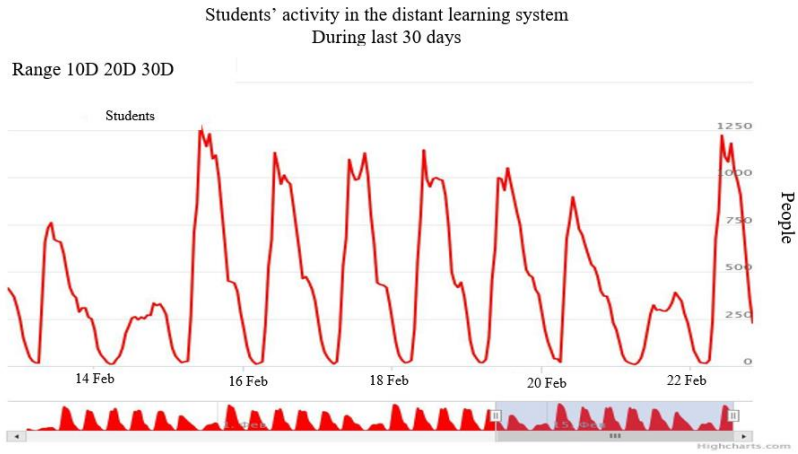


Figure 11
Students' activity in the system

As you can see from the graphs of activity, up to 1500 users are simultaneously daily in the system.

To assess the performance of the server part of the system, we will analyze the operation of the system's web server. To do this, we will use the log of the Microsoft Internet Information Services 10.0 web server for the period from December 1 to December 31, 2020. Summary data on the analysis of the log are given in Table 1. This table contains the following information:

- Request handler – is a server-side script that processes the user's request
- Total requests – the total number of requests to this handler for the period under review.
- Bytes / request received – the average value of received bytes per 1 request
- Bytes / Request Sent – the average value of sent bytes per 1 request

Average execution time (ms) – the average execution time of 1 request in milliseconds

As you can see from the table, most queries are executed within 500 ms, which indicates the performance of the system since it is considered that if the user's query lasts less than 1 second, then he believes that the system is working without delays.

We can also see that there are requests that take more than 1 second to complete. These requests are related to uploading files or downloading files. The speed of these requests directly depends on the volume of transmitted data and, accordingly, increases with a large amount of data. Users are aware of this file and do not perceive this phenomenon as a lack of system response.

Table 1

System web server log analysis data

Request handler	**Total requests**	**Byte / request received**	**Byte / request sent**	**Average execution time (ms)**
/SDO/Chat/Handler (Chat operation)	34921729	2805	477	294
/SDO/Service (user search)	4383143	25395	810	360
/SDO/Course/Details.aspx (course page)	1031793	2826	14062	679
/SDO/Test/ImageTest.aspx (picture from the base for the test)	761530	2412	7588	458
/SDO/Chat/Default.aspx (Course chat page)	500220	2215	6662	402
/SDO/Default.aspx (List of courses for the user)	440763	2098	5673	1107
/SDO/Entry.aspx (Home page)	398081	1743	21641	604
/SDO/Course/Score.aspx (Course performance)	220684	2320	9432	395
/SDO/Messages.aspx (Private messages)	207699	3842	8445	567
/SDO/Service/Work-study (Submitting / uploading completed work on the course)	139504	2338	1001810	4058
/SDO/Service/Work (uploading a file with an assignment)	118592	2124	474271	4166
/SDO/Course/Work.aspx (assessment of completed assignments)	112530	18094	18661	541
/SDO/Service/Resource (loading course resource)	105809	2184	652169	3678
/SDO/Test/Detail.aspx (view loaded test)	65531	2001	3780	272
/SDO/Test/TestForm.aspx (test)	48088	2854	5342	814
/SDO/Course/AddWork.aspx (adding an assignment to a course)	36834	7526	8562	337
/SDO/Course/ActiveWorks.aspx (list of unchecked assignments)	28709	3316	12812	1389
/SDO/Service/PrivMessFile (private messages file)	19905	2525	1232567	3878
/SDO/Course/Group.aspx (group list of the course)	17417	2491	8847	465

/SDO/Test/LoadTest.aspx (loading test for course)	13958	7050	13739	396
/SDO/Course/AddResource.aspx (adding a resource to a course)	11117	5796	5641	326
/SDO/Course/GraphComm.aspx (schedule of assignments for the course)	10322	2451	8166	334
/SDO/Group.aspx (group list)	9372	2413	6377	417
/SDO/Forum/Default.aspx (course forum)	7220	1963	6650	408
/SDO/Test/Appelation.aspx (test appellation)	6550	4746	6189	372
/SDO/Forum/Topic.aspx (course form topic)	4497	4022	9374	496
/SDO/Instruction.aspx (system instructions)	4021	2148	4036	824
/SDO/Test/TestResult.aspx (test result)	3691	3276	8667	621
/SDO/Course/Log.aspx (activity log)	3010	13823	8506	501
/SDO/Test/ViewTest.aspx (view the completed test)	1813	2882	19127	362

Based on the system operation's data, we will assess the system performance based on the indicators from paragraph 1.1.

Based on the given data, it has the following initial values:

- Intensity of requests to the system $\lambda = 978$ requests / minute
- Average time to service a request t_rst = 361 ms = 0.006 minutes
- Number of service channels n = 10,000 (the e-learning system web application is configured for this number of simultaneous processing)
- Average size of a response to a request in kilobytes - 7.89 Kbytes

Then we calculate the performance indicators of the system:

- Intensity of requests $\rho = \lambda * t\_rst = 978 * 0.006 = 5.87$, i.e. on average, the system simultaneously processes about 6 requests.
- Probability of failure - as described in Section 1.1, with a large value of the number of service channels and intensity, the probability of failure actually becomes equal to 0.
- Nominal performance of the system $N = n / \rho = 10000 / 5.87 = 1704.2$, i.e. on average, the system can process about 1,700 applications per minute.
- Actual performance of the system $F = \lambda / N = 978 / 1704.2 = 0.574$, i.e. on average, the system is loaded by 57%.

- The required size of the data transmission channel:

$S = (\lambda * r) / (125 * 60) = (978 * 7.89) / (125 * 60) = 1.03$ Mbps

The graph (Fig. 12) shows the load on the network component of the web server, which shows the average input and output bytes per second.

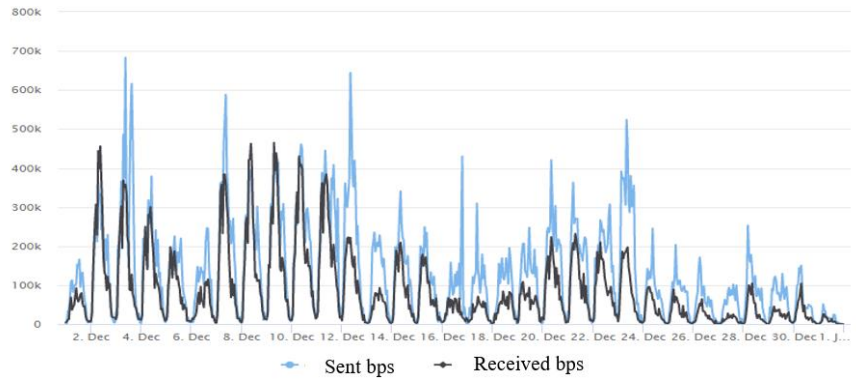


Figure 12

Load on the network component of the web server

As you can see from the presented graph, the traffic generated by the web server can easily be served by the network interfaces on the available resources.

Conclusions

Currently, there are many e-learning systems that universities use in the educational process. Such systems are often included in portal solutions (educational portals) that are available in universities and provide a single-entry point for different categories of users.

In this study, the main hardware and software components of the e-learning system were identified as part of the University's Educational Portal.

The model for assessing the functioning of the e-learning system presented in the article used the QS criteria: performance analysis (system response time to user requests) and system reliability (no failures when working with the system).

VKTU named after D. Serikbayev has its own portal solution (www.do.ektu.kz), with an e-learning system that allows implementation of e-learning, combining all participants in the educational process and is an integral part of the unified educational environment of the University. The implementation of the e-learning system within a single portal solution, made it possible to achieve the high speed of the e-learning system, by reducing the overhead costs associated with the coordination of the work of heterogeneous systems for managing the educational process at the University. The system allowed for improvements, the quality of work and interaction of various users in the system and, as the analysis of the work of this system showed, it is quite effective.

Acknowledgment

This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan under the grant financing program for the 2020-2022 years by the program 08856846 “Methodology for creating a liberal model of On-Line education for higher education institutions of the Republic of Kazakhstan”.

References

[1] T. Y. Wang, C. H. Wang, «E-Learning Platform of STEAM Aesthetic Course Materials Based on User Experience» In Proc. of the 1st International Cognitive Cities Conference (IC3), IEEE, August 2018, pp. 123-128

[2] P. Zaharias, C. Pappas. Quality management of learning management systems: A user experience perspective, in Current Issues in Emerging eLearning, Vol. 3(1), 2016, p. 5

[3] R. Kraleva. Designing an Interface for a Mobile Application Based on Children's Opinion, in International Journal of Interactive Mobile Technologies, Vol. 11(1), 2017, pp. 53-70

[4] Legkov K. E. Models of information subsystems of automated control systems for complex objects, in T-Comm. 2017, No. 5

[5] J. W. Lin, H. C. K. Lin. User acceptance in a computer supported collaborative learning (CSCL) environment with social network awareness (SNA) support, in Australasian Journal of Educational Technology, Vol. 35(1), 2019, pp. 100-115

[6] V. Narayan, J. Herrington, and T. Cochrane. Design principles for heutagogical learning: Implementing student-determined learning with mobile and social media tools, in Australasian Journal of Educational Technology, Vol. 35(3), 2019, pp. 86-101

[7] I. Valova, M. Marinov. Facebook as a Tool Aiding University Education-Whether it is Possible and Useful, in TEM Journal, Vol. 8(2), 2019, pp. 670-676

[8] Dolzhenko A. I. Modeling of a corporate information system, in Bulletin of higher educational institutions. North Caucasian region. Series: Social Sciences. 2006, No. 2

[9] Bain A. M. Mathematical model of the distance learning system, in Bulletin of higher educational institutions. Electronics, 2010, No. 2 (82), pp. 71-75

[10] Mutanov G. M., Shakarimova A. B. Educational portal of the university. Theory and practice. - Ust-Kamenogorsk: EKSTU, 2006, 352 p.

[11] Transformation of a technical university into an innovative university: methodology and practice / Ed. G. M. Mutanov. - Ust-Kamenogorsk: EKSTU, 2007, 480 p.

[12] Dmitrenko T. A., Derkach T. N., Dmitrenko A. A. Technology of developing a distance learning system, in Economics. Informatics. 2014, No. 8-1 (179) pp. 128-137

[13] Naumov A. V., Mkhitaryan G. A., Rybalko A. A. Software set of intellectual support and security of lms mai class. Net, in Bulletin of SUSU. Series: Mathematical Modeling and Programming. 2016. No. 4, pp. 129-140

[14] Kumargazhanova, S., Baklanov, A., Györök, Gy., et al.: Development of the Information and Analytical System in the Control of Management of University Scientific and Educational Activities, Acta Polytechnica Hungarica, Vol. 15, No. 4, 2018, pp. 27-44, DOI: 10.12700/APH.15.4.2018.4.2

[15] Baklanov, A, Baklanova, O, Grigoryeva, S, Kumargazhanova, S, Györök, G, et al.: The Development of Hybrid IP Architecture for Solving the Problems of Heating Networks, in Acta Polytechnica Hungarica, №1 (17), 2020, pp. 123-140, DOI: 10.12700/APH.17.1.2020.1.7

[16] Rybakova, D., Sygynganova, I., Kumargazhanova, S., Baklanov, A., Shvets, O.: Application of a CPU Streaming Technology to Work of the Computer with Data Coming from the Network on the Example of a Heating Station. 18th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM), Erlagol, Russia, 2017, pp. 128-130

[17] Kassymkhanova, D., Kurochkin, D., Denissova, N., Kumargazhanova, S., Tlebaldinova, A. Majority voting approach and fuzzy logic rules in license plate recognition process, in The 8th International Conference on Application of Information and Communication Technologies (AICT 2014), Astana, 2014, pp. 155-159

[18] Karymsakova, I., Denissova, N., Kumargazhanova, S., Krak, I. Robotic plasma spraying system for implants of complex structure: 3d model and motion planning, in International Journal of Computing, 19(2) 2020, pp. 224-232. On-line ISSN 2312-5381

[19] Azarova, A., Azarova, L., Pavlov, S., Savina, N., Kaplun, I., Wójcik, W., Smailova, S., Kalizhanova, A. Information technologies for assessing the quality of it-specialties graduates' training of university by means of fuzzy logic and neural networks, in International Journal of Electronics and Telecommunications, 66(3) 2020, pp. 411-416

[20] Uvaliyeva, I., Smailova, S. Development of decision support system to control the quality of education, in Proceedings of the International Conference «Application of Information and Communication Technologies-AICT 2014», Astana, 2014, pp. 528-532

[21] Lutsiv, N., Maksymyuk, T., Beshley, M., Lavriv, O., Vokorokos, L., Gazda, J. (2022) Deep Semisupervised Learning-Based Network Anomaly Detection in Heterogeneous Information Systems. CMC-Computers, Materials & Continua, 70(1) 413-431

Determination of Critical Deformation Regions of a Lithium Polymer Battery by DIC Measurement and WOWA Filter

Szabolcs Kocsis Szürke[1], Adrienn Dineva[1], Szabolcs Szalai[2], István Lakatos[3]

[1] Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary, kocsis.szabolcs@ga.sze.hu, dineva.adrienn@sze.hu

[2] Department of Vehicle Manufacturing, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary, szalaisz@sze.hu

[3] Department of Automotive and Railway Engineering, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary, lakatos@sze.hu

Abstract: This paper considers the determination method of deformation location of lithium polymer batteries. Measurements are performed using the Digital Image Correlation (DIC) technique and the obtained results are sorted into a database as a function of the charge level. A statistically based algorithm is used to eliminate measurement errors and outliers. This paper adopts the Weighted Ordered Weighted Averaging (WOWA) operator-based 2D filtering method with the purpose of determining the critical regions of the cell. During the tests, several lithium polymer batteries of the same type but in different states are compared. Measurements on completely new and also on worn-out batteries are performed. The results support that the regions where greater deformation is expected during charging and discharging can be predicted. Results of investigations validate that the proposed approach is suitable for determining the critical deformation regions with high accuracy.

Keywords: battery swelling; battery testing; lithium polymer battery; GOM Atos; DIC measurement; WOWA

1 Introduction

Lithium batteries are currently widely used as a popular energy storage device in the automotive industry and among portable electronic devices as well. The main arguments in favor of them are high energy density and long stable operation. In the past years, lithium-based batteries have been getting widespread because of the increasing demand, financial investments and technological advantages.

However, there are several developments and operational issues with their use, such as minimizing these faults, which is important for users. Numerous studies have also dealt with the state of charge (SOC) and state of health (SOH) estimation [1-4], temperature effect [5], or examination of cells at different pressures [43-45]. In addition to traditional tests, mechanical-based measurements have recently become popular, supplementing them to provide more accurate information on the internal state of lithium-ion batteries [6]. Studies have also been performed on cell impact [7-8], mechanical deformation [9] and pressure [10]. In addition to external effects, improper operation, overcharging [11], deep discharge, and a high number of cycles can lead to deformation and swelling of the cells. The main causes of this may be an expansion of the host materials, an expansion in the volume of the electrode, a change in pressure in the dead space of the cell, or gas formation [12-14]. In addition to continuous use, the number of cycles also increases, in which reversible volume change can also become an irreversible process. This involves a mechanical reaction of the battery cells, which can cause loss of capacity and failure. For a deeper analysis of structural properties, the reader is referred to [13] [15] [16] papers for additional details. Furthermore, continuous deformation can be observed even under normal, manufacturer-recommended use. In accordance with the State-of-the-Art, determining the critical location of this deformation is in the focus of this paper. Well-defined cell parameters and diameters are important factors in planning battery placement as well as SOH estimation. For this reason, measuring the deformation of lithium batteries has become a popular area of research in recent times. Several methods are used to measure deformation during use. Tactile tests measure the deformation of a cell at one or a few points, [17-21] it can even be tested together with the effect of pressure [22]. In most cases, a displacement sensor is used to measure them [23], placed in the middle of the battery, on this basis, a deformation map is even made [24]. The advantage of this method is that they can even examine the cells at the system level [25] and the deformation during storage continuously measurable [26]. The problem with this is that the battery varies asymmetrically and amorphously, making it difficult to determine where to measure. In addition, a number of good and useful publications have been produced with other sensors and methods of deformation measurement: thickness gauge [27], ultrasonic transducer pulser and receiver [28], strain sensor [29] or high resolution dilatometry [30]. The obtained data in force-SOC combination can even be used for charge-level estimation [31]. To better understand cells without destruction, CT scans are performed to look for structural defects [32] or to study the structural change of bad cells [33]. A detailed analysis of the methods used to measure deformation can be found in the following publications [34-35]. In this publication, we used the popular DIC technique for high-precision analysis of deformation. Thanks to its many benefits - such as easy experimental setups, simple implementation, high resistance to environmental influences, variability and widely adjustable time and space resolution - DIC technique has become widely accepted as a powerful and flexible tool for

measuring the movement and deformation of different materials. With this solution, measurement accuracies of up to +/– 1-2 µm can be achieved. The GOM system has been used successfully in a number of areas [36-37]. It is also used in the field of batteries, mostly for electrode composition testing [38] and structural characteristics measurement [39] but has also been used for deformation measurement [40-41] possibly in combination with other methods [42]. This paper is structured according to the following: The measurement process section describes the measurement procedure, tools, and data storage. In the third chapter, the adoption of the WOWA operator is presented. The evaluation and results section presents the measurement results using a filtering procedure.

2 Measurement Process

Intercalation between lithium batteries occurs during charging and discharging, not in no-load condition or a discharged state, thus, there is no or a very low degree of deformation during the interruption. This assumption was confirmed in our previous publication by results from displacement sensor measurements [46]. Consequently, tests can be interrupted and optical measurements can be made. In order to achieve greater accuracy and better mapping of critical locations, the tests were performed using DIC technique. The meaning of a DIC technique is Digital Image Correlation. According to the literature with this technique can be measure displacement, deformation, 3D coordinates and can be made 3D scanning. In this article, a 3D coordinates measurement was used with a 3D scanning DIC system. The DIC system was the GOM Atos TripleScan II hardware with GOM Atos Professional software. The ATOS Triple Scan non-contact structured blue light 3D scanner is a type of coordinate measuring machine that measures millions of points per single scan/measurement. It uses advanced measuring and projection techniques to produce high quality data and precision accuracy for full-object dimensional analysis. ATOS sensors are self-monitoring systems. The sensors identify changing ambient conditions during operation. The software of the sensors is continuously monitoring the calibration status, the transformation accuracy as well as environmental changes and part movements in order to ensure the quality of the measuring data. In this research the battery was measured several times and after that the individual measurements were assembled in the software. This allowed a detailed analysis of the deformation. Several conditions must be fulfilled for accurate measurement, therefore, the DIC technique requires the following:

- Reduce the reflection: The quality of preparation is important for DIC measurements. Thus, the object to be measured should not be reflective, otherwise, the test element should be thinly coated with special anti-reflex paint [47]. In this article, a MR2000 anti-reflex spray was used. (In our case, the surface had to be treated.)

- Use reference points: Another important thing is that the camera detects at least three reference points during the measurement. The GOM Atos system uses special coded reference points in several sizes. For the measurements was used 1mm and 0.8 mm coded reference points.
- Taking pictures: In these tests, we used the two-camera GOM ATOS measuring unit, where one camera digitizes based on the reference points and the other functions as a control.
- Image analysis: After recording, the deformation images are compared to a reference or initial image using a special cross-correlation algorithm that will extract the displacement fields.

During the tests, the following devices were used to determine the electrical parameters: the power supply - Hameg HMP 4030, the dummy load - EL3000, the data acquisition card - NI 9201 and lithium polymer battery – Turnigy nano-tech LiPo 5 Ah. LabView software is responsible for test control. The tests were performed on 3 different 5 Ah lithium polymer batteries. The cells were of the same in type, size and capacity, the only difference being the production time. A new and old battery was used during the measurements. In the following distribution:

- Battery number 1 (approximately 80% capacity): charging 4.2 V and discharging up to 3 V.
- Battery number 1 (6 months later):
    - Charging 4.2 V and discharging up to 3 V.
    - Charging 4.2 V and discharging up to 0 V.
- Battery number 2 (approximately 30% capacity): charging 4.2 V and discharging up to 0 V.
- Battery number 3 (approximately 100% capacity): charging 4.2 V and discharging up to 3 V.

A total of 5 charge and 5 discharge tests were performed. The general course of the tests was as follows: fully charge the battery 100% -SOC; digitization of a fully charged state; start discharging and interrupt, digitize every 360 s; in general, surface measurements are made: 100%, 90%… 10% and 0% of the charge level; conditioning for one hour; re-digitization of a fully discharged state; start charging and interrupt, digitize every 360 s. In general, surface measurements are made at 0%, 10%… 90% and 100% of the charge level. After battery replacement, these steps were performed on each cell.

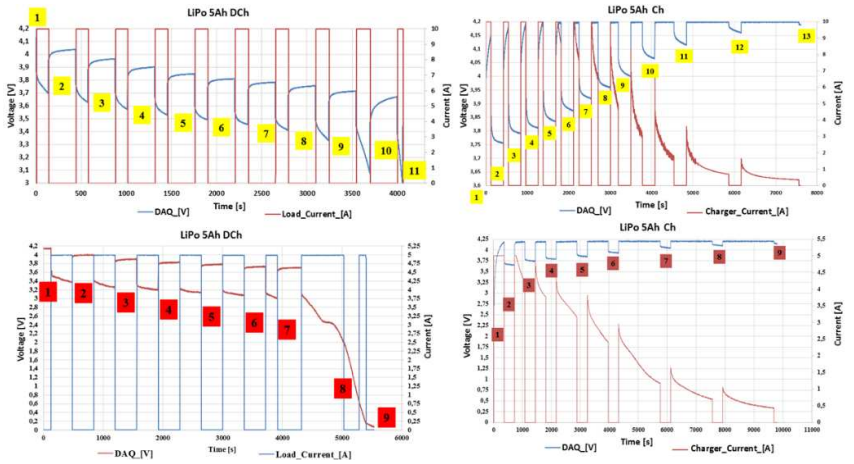


Figure 1
Charging and discharging characteristics of LiPo batteries

Figure 1 shows the charging and discharging current and voltage profile of a normal (3 V) and a fully discharged (0 V) cell. The upper part of the Fig. 1 shows the voltage / current diagrams obtained during the first test, from the left during discharge and from the right during charging. The bottom of the Fig. 1 shows the results of cell number two, in this test we discharged the battery to 0 V. In this case, too, the voltage / current profiles measured during discharge are shown on the left and the charging on the right. The numbering shown in the Fig. 1 can be observed at interruptions, digitization has taken place at these points. Figure 1 shows that the number of interrupts varied as a function of the length of the tests. For example, the upper right Fig. 1 had 9 interrupts, representing 11 images (including the start and end states), and the lower left figure had 7 interrupts, representing 9 images. The reason for this is that the battery discharges faster. The number of digitization points created during charging also varies, which are due to the faster charging of the cell or optimization of the measurement time. In some cases, at the end of the charge, we digitized every 10 to 20 minutes because in this case the charged energy is less, so the degree of deformation is smaller based on the observations. Furthermore, in each case we store the data depending on the charge level, therefore, it will not cause an issue with a different number of digitization points. Figure 2 shows the results of four measurements, in each case also showing the A (on the right side of the picture pair) and B (on the left side of the picture pair) side of the battery. The cells change asymmetrically and amorphously. In general, they swell during charging and contract during discharge. Figure 2 shows the largest deviation from the initial state, which is 100% SOC during charging and 0% SOC during discharge. The lower left corner shows the results during charging, and the lower right corner shows the fully discharged state up to 0 V. Based on these few deep discharges, it was observed that in the range below 3 V, contraction is replaced by swelling.

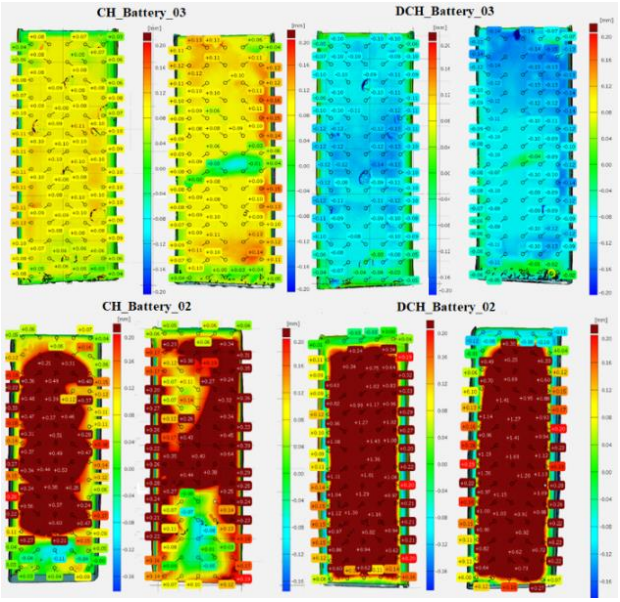
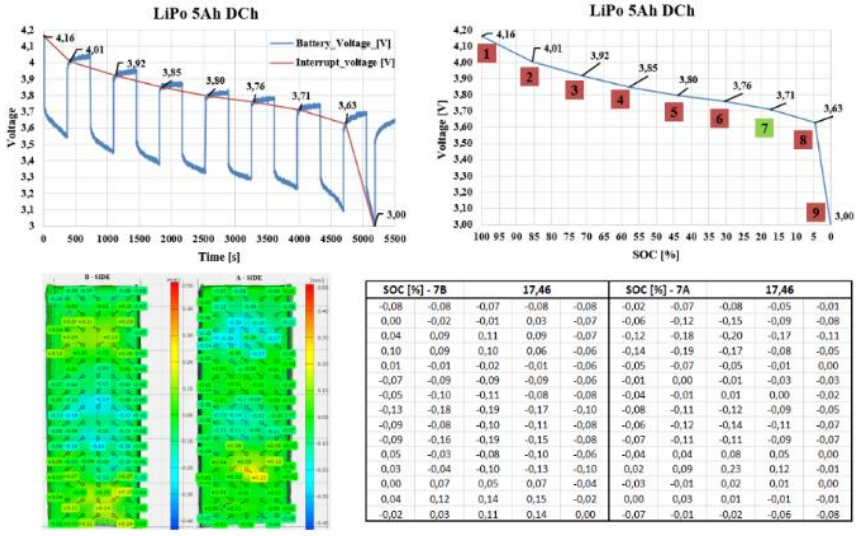


Figure 2
Deformation images recorded using the DIC technique

All electrical and surface digitization points are stored together and arranged in data matrices depending on the charge level.



SOC [%] - 7B			17,46		SOC [%] - 7A			17,46	
-0,08	-0,08	-0,07	-0,08	-0,08	-0,02	-0,07	-0,08	-0,05	-0,01
0,00	-0,02	-0,01	0,03	-0,07	-0,06	-0,12	-0,15	-0,09	-0,08
0,04	0,09	0,11	0,09	-0,07	-0,12	-0,18	-0,20	-0,17	-0,11
0,10	0,09	0,10	0,06	-0,06	-0,14	-0,19	-0,17	-0,08	-0,05
0,01	-0,01	-0,02	-0,01	-0,06	-0,05	-0,07	-0,05	-0,01	0,00
-0,07	-0,09	-0,09	-0,09	-0,06	-0,01	0,00	-0,01	-0,03	-0,03
-0,05	-0,10	-0,11	-0,08	-0,08	-0,04	-0,01	0,01	0,00	-0,02
-0,13	-0,18	-0,19	-0,17	-0,10	-0,08	-0,11	-0,12	-0,09	-0,05
-0,09	-0,08	-0,10	-0,11	-0,08	-0,06	-0,12	-0,14	-0,11	-0,07
-0,09	-0,16	-0,19	-0,15	-0,08	-0,07	-0,11	-0,11	-0,09	-0,07
0,05	-0,03	-0,08	-0,10	-0,06	-0,04	0,04	0,08	0,05	0,00
0,03	-0,04	-0,10	-0,13	-0,10	0,02	0,09	0,23	0,12	-0,01
0,00	0,07	0,05	0,07	-0,04	-0,03	-0,01	0,02	0,01	0,00
0,04	0,12	0,14	0,15	-0,02	0,00	0,03	0,01	-0,01	-0,01
-0,02	0,03	0,11	0,14	0,00	-0,07	-0,01	-0,02	-0,06	-0,08

Figure 3
Storing surface digitization data in a database

Figure 3 shows the data aggregation. The results during the discharge of cell 2 can be seen in the upper left image of the Fig. 3, in addition to the voltage diagram, the points of the interruptions are also shown. In the upper right image the same points are shown, but already as a function of the charge level. The Coulomb Counting method was used to determine the charge level. The digitization point 7 is marked in green on the diagram, which is shown in detail in the lower part of Fig. 3. The lower left image displays the digitization data and the lower right image shows the data matrix as a function of the charge level.

3 Adoption of WOWA Operator for Critical Region Detection

From the millions of measurement points created during digitization, we selected 75-80 measurement sites evenly spaced and stored them in the data matrix. The recording of equally spaced points is automatically networked and recorded by the GOM measuring system software. It is important to note that the same measurement uses the same mesh for all charge levels. There is a relatively short time (5 minutes) available for digitization during discharge and charge interruptions, so in some cases, errors may occur. Therefore, smaller holes are created where there is no accurate deformation information and incorrect data can be entered here. Therefore, the results should be evaluated by a statistical method in order to effectively filter the database from errors and outliers. Several methods are used in the literature for similar problems [48-51]. Torres et al. present an efficient method for post-filtering images using the WOWA-operator [52], whose applicability in critical region determination is considered in this paper.

3.1 Mathematical Background of Aggregation Operators

The need of information fusion has become increasingly important in various disciplines of modern engineering and artificial intelligence [53]. The aggregation functions are mathematical functions that are used to incorporate various information. The arithmetic mean and the weighted mean are the most well-known aggregation operators. The main characteristic of the weighted mean is that it permits us to weight the different data according to their relevance that is not possible with the arithmetic mean. The arithmetic mean is given in Eq. (1).

$$W^{AM}(a_1, \ldots, a_n) = \sum_{k=1}^{n} \frac{a_k}{n} \tag{1}$$

The weighted mean can be formulated as follows:

$$W^{WM}(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i a_i \tag{2}$$

where w_i stand for the weights of the k^{th} data source. The weights are positive and $\sum w_i = 1$. The classical aggregation operators such the weighted average, are key tools of decision theory [54]. However, nowadays they are considered particular families of more general aggregation operators. Originally, the possible aggregation of fuzzy sets required operations that provide a single fuzzy number as a result of combining several fuzzy numbers [55]. Many fuzzy aggregation operators have been presented, such as the product, t-norms, different types of means, etc. The Ordered Weighted Averaging (OWA) operator has been introduced in [56] by Yager, that is a class of generalized mean operator. The OWA operator has the form [56]:

$$W^{OWA}(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i a_{\sigma(i)} \tag{3}$$

in which σ(i) corresponds to the permutation of a_i, i.e. the elements of the data vector are sorted decreasing order, from the largest value to the lowest one. This sorting allows to give the desired importance to the largest, lowest or medium value data. Several papers have been published about the properties and modifications of aggregation operators [57]. It is worth noting that OWA aggregation functions and weighted arithmetic means are special cases of Choquet integral. A large family of aggregation functions based on Choquet integrals [58]. Their detailed properties and definition are given with respect to a fuzzy measure [57]. When the fuzzy measure is additive, Choquet integrals become weighted arithmetic means, and when the fuzzy measure is symmetric, they become OWA functions. Different OWA operators are distinguished in the literature according to their weighting function. In [59] Torra has introduced the Weighted OWA (WOWA) aggregation function, that can be considered as the generalization of weighted means. The weighted OWA function has two sets of weights, one of them has the same function as the weighting vector in weighted means, wherease the other is equivalent to the weighting vector in OWA.

3.2 Outline of the Proposed Method

Due to its excellent properties the WOWA operator has wide application possibilities. For instance, image processing techniques can benefit from such an approach. The efficiency of this method is that two vectors can be used to well set the importance of data and filter out erroneous data. The other thing that makes a statistically based analysis useful is that the differences between points close to each other can be well filtered. The goal with these measurements is to determine

a critical region, so outliers can make this very difficult. The first step in processing the data is to read the database and rearrange it using the algorithm:

SOC [%] - 9B		0,00		
-0,04	-0,10	-0,12	-0,07	-0,03
-0,09	-0,18	-0,19	-0,13	-0,11
-0,16	-0,23	-0,25	-0,22	-0,15
-0,18	-0,24	-0,24	-0,14	-0,07
-0,08	-0,11	-0,10	-0,04	-0,02
-0,05	-0,02	-0,04	-0,06	-0,05
-0,06	-0,03	-0,02	-0,03	-0,03
-0,11	-0,14	-0,16	-0,12	-0,07
-0,09	-0,17	-0,20	-0,15	-0,09
-0,10	-0,16	-0,16	-0,14	-0,10
-0,07	0,00	0,04	0,02	-0,02
0,00	0,05	0,20	0,09	-0,02
-0,06	-0,06	-0,03	-0,03	-0,02
-0,01	0,00	-0,02	-0,03	-0,02
-0,10	-0,03	-0,05	-0,08	-0,10

Figure 4
Rearrange the database

In the figure, the point to be determined is marked in yellow, and the data of the 3 X 3 matrix marked in blue were used for the determination. Based on this, a total of 9 values were used to determine a selected point (marked in yellow). The following relationship describes the coordinate points of the data extracted from the matrix:

$$\begin{matrix} \mathrm{I}(\mathrm{x}-1;\,\mathrm{y}-1) & \mathrm{I}(\mathrm{x}-1;\,\mathrm{y}) & \mathrm{I}(\mathrm{x}-1;\,\mathrm{y}+1) \\ \mathrm{I}(\mathrm{x};\,\mathrm{y}-1) & \mathrm{I}(\mathrm{x};\,\mathrm{y}) & \mathrm{I}(\mathrm{x};\,\mathrm{y}+1) \\ \mathrm{I}(\mathrm{x}+1;\,\mathrm{y}-1) & \mathrm{I}(\mathrm{x}+1;\,\mathrm{y}) & \mathrm{I}(\mathrm{x}+1;\,\mathrm{y}+1) \end{matrix} \quad \textbf{(4)}$$

The data is transferred from the matrix to the row vector according to the following formula:

$$\begin{matrix} I_3 & I_6 & I_9 \\ I_2 & I_5 & I_8 \\ I_1 & I_4 & I_7 \end{matrix} = [I_1\, I_2\, I_3\, I_4\, I_5\, I_6\, I_7\, I_8\, I_9\,] = a \quad \textbf{(5)}$$

When determining weight vectors, all values were considered to be of equal significance.

$$p = [1/9\,;1/9\,;1/9;1/9;\,1/9;\,1/9;1/9;\,1/9;\,1/9] \quad \textbf{(6)}$$

An empirically determined weight vector can be further investigated if even finer critical region detection performance is required by emphasizing distinct parts of the battery. When specifying the second vector, our purpose is to eliminate, filter out outliers and highlight critical regions.

$$W_n = [0;\,0{,}075;\,0{,}125;\,0{,}175;\,0{,}250;\,0{,}175;\,0{,}125;\,0{,}075;\,0] \quad (7)$$

As a first step, we determined the coordinate points based on the following formula:

$$\left(i/n; \sum_{j \le i} \omega_j\right) \tag{8}$$

$$i = 1 \quad (1/9;\ \omega_1) \quad (1/9; 0)$$

$$i = 2 \quad (2/9;\ \omega_1 + \omega_2) \quad (2/9; 0{,}075)$$

$$i = 3 \quad (3/9;\ \omega_1 + \omega_2 + \omega_3) \quad (1/3; 0{,}2)$$

$$i = 4 \quad (4/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4) \quad (4/9; 0{,}375)$$

$$i = 5 \quad (5/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5) \quad (5/9; 0{,}625)$$

$$i = 6 \quad (6/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6) \quad (2/3; 0{,}8)$$

$$i = 7 \quad (7/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 + \omega_7) \quad (7/9; 0{,}925)$$

$$i = 8 \quad (8/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 + \omega_7 + \omega_8) \quad (8/9; 1)$$

$$i = 9 \quad (9/9;\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 + \omega_7 + \omega_8 + \omega_9) \quad (1; 1)$$

After curve fitting, the following function was obtained:

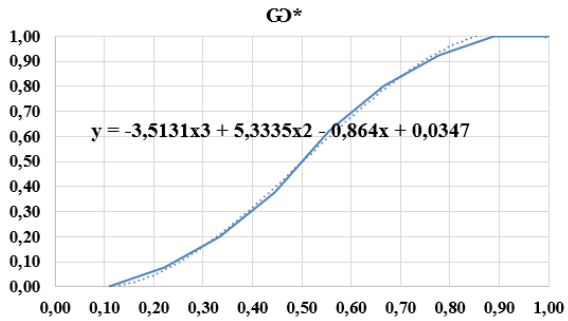


Figure 5
Determined ω* after curve fitting

The equation after fitting the curve is as follows:

$$\omega^*(x) = -3{,}513x^3 + 5{,}3335x^2 - 0{,}864x + 0{,}0347 \tag{9}$$

Using ω*(x) we determined the weights:

$$i = 1 \;\; \omega_1 = \omega^*(p_1) \cong 0$$

$$i = 2 \;\; \omega_2 = \omega^*\left(\sum_{i=1}^{2} p_i\right) - \omega^*(p_1) \cong 0{,}068$$

$$i = 3 \;\; \omega_3 = \omega^*\left(\sum_{i=1}^{3} p_i\right) - \omega^*\left(\sum_{i=1}^{2} p_i\right) \cong 0{,}142$$

$$i = 4 \;\; \omega_4 = \omega^*\left(\sum_{i=1}^{4} p_i\right) - \omega^*\left(\sum_{i=1}^{3} p_i\right) \cong 0{,}187$$

$$i = 5 \;\; \omega_5 = \omega^*\left(\sum_{i=1}^{5} p_i\right) - \omega^*\left(\sum_{i=1}^{4} p_i\right) \cong 0{,}206$$

$$i = 6 \;\; \omega_6 = \omega^*\left(\sum_{i=1}^{6} p_i\right) - \omega^*\left(\sum_{i=1}^{5} p_i\right) \cong 0{,}187$$

$$i = 7 \;\; \omega_7 = \omega^*\left(\sum_{i=1}^{7} p_i\right) - \omega^*\left(\sum_{i=1}^{6} p_i\right) \cong 0{,}142$$

$$i = 8 \;\; \omega_8 = \omega^*\left(\sum_{i=1}^{8} p_i\right) - \omega^*\left(\sum_{i=1}^{7} p_i\right) \cong 0{,}068$$

$$i = 9 \;\; \omega_9 = \omega^*\left(\sum_{i=1}^{9} p_i\right) - \omega^*\left(\sum_{i=1}^{8} p_i\right) \cong 0$$

The obtained weight vector is as follows:

$$\omega = [0; 0{,}068; 0{,}142; 0{,}187; 0{,}206; 0{,}187; 0{,}142; 0{,}068; 0] \tag{10}$$

The first step in evaluating the values was to load the data (heatmap of deformation [mm]):

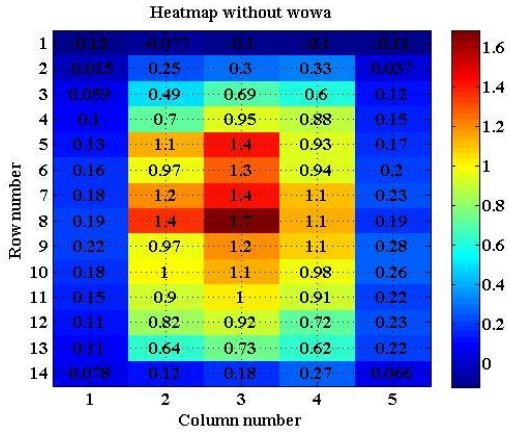


Figure 6

The last point of discharge is without the WOWA- based method

The second step is to determine the corresponding 3X3 matrix for each data point based on Eq. (4). The determined values were then sorted in ascending order and applied to the following summary:

$$f_{WOWA} = \sum_{i=1}^{9} \omega_i * a_i \tag{11}$$

The final deformation point matrix was as follows:

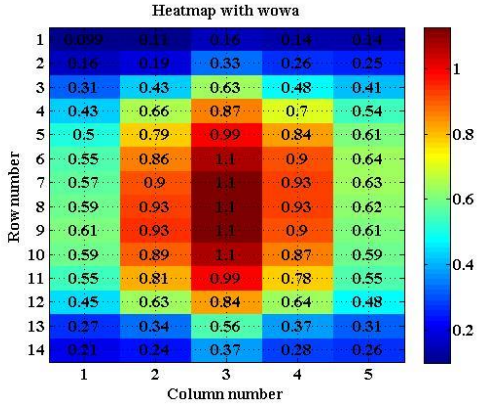


Figure 7

The last point of discharge after using the WOWA- based method

The values shown in Fig. 7 the dimension is [mm], the battery is divided into 5 columns and 14 rows. During the test, the battery was discharged to 0 V. Based on the results, it can be concluded that this type of deep discharge causes significant swelling. Based on the heat map, it can be observed that the middle part of the cell changed the most. It can be observed that the filtered data eliminate outliers, possible errors and form a more uniform heat map. We also visualized the results for better presentation. Figure 8 shows a comparison of the measured and filtered values after visualization.

Visualization allows to better observe how the algorithm equalizes the surface of the battery. By eliminating the local minimum and maximum, the critical zone can be more easily selected. During this test, the greatest deformation occurred in the middle of the cell. More specifically, points 6-7-8-9-10 in column 3 are the most critical.

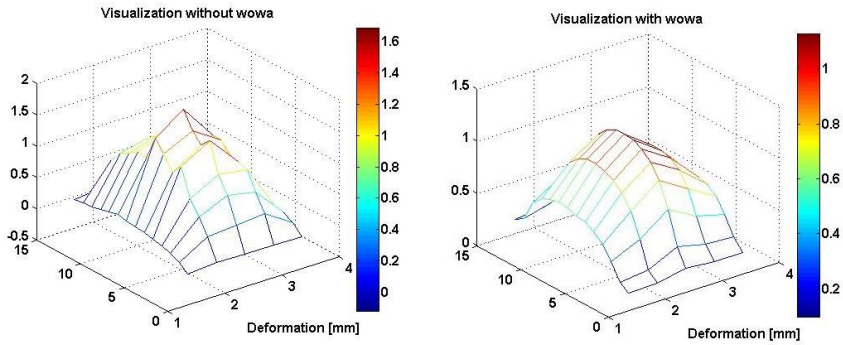


Figure 8
Visualization of the results obtained with the WOWA- based method

4 Evaluation and Results

Using the WOWA-based method, we filtered all states from outliers and potentially erroneous data and created a unified deformation map. To determine the critical regions, all digitization states and measured points of each measurement had to be examined. During the analysis, we observed that the examined point was smaller or larger than the average deformation point, based on this, it was classified as 1 or 0. The average deformation point, in this case, is the average of the data on the unchanged side of the cell measured at the same charge level. In all cases, the value 1 represents the critical deformation greater than average. Values higher than average during charging and lower values during discharge were considered 1 because contraction is expected during discharge. Figure 9 shows the results obtained when charging side B of battery number 2 in the case of 100% SOC.

The upper part of the Fig. 9 shows the analyzed data, the left image shows the original data, and the right image shows the filtered results. The bottom of the Fig. 9 shows the results of the under/over test, the original data is shown in the left image (mean value 0.109 mm) and the results filtered by the WOWA algorithm are shown in the right image (mean value 0.108 mm). The next step in the analysis was to fit the data to each other, hence to summarize the critical points of the measurements for the same measurement and side (separately for A and B).

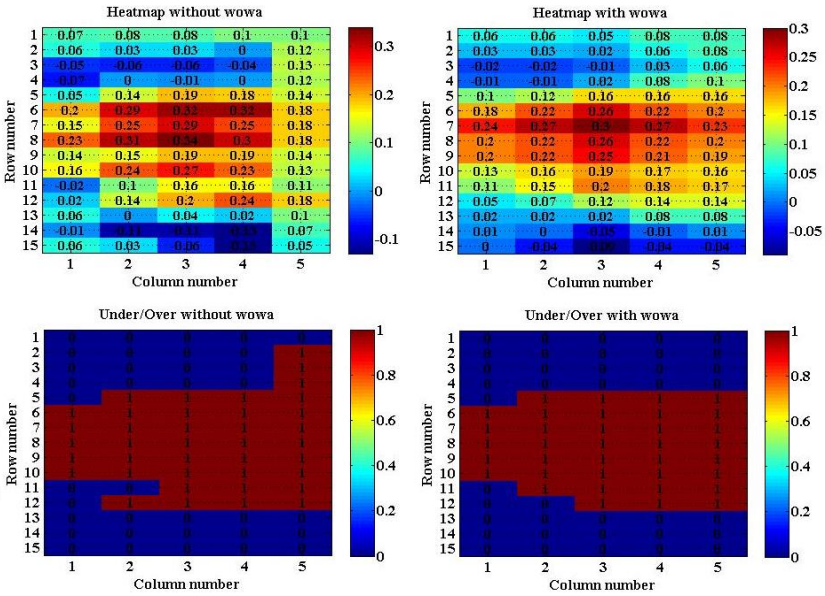


Figure 9
Deformation average test

In Fig. 10 shows the results obtained after data alignment:

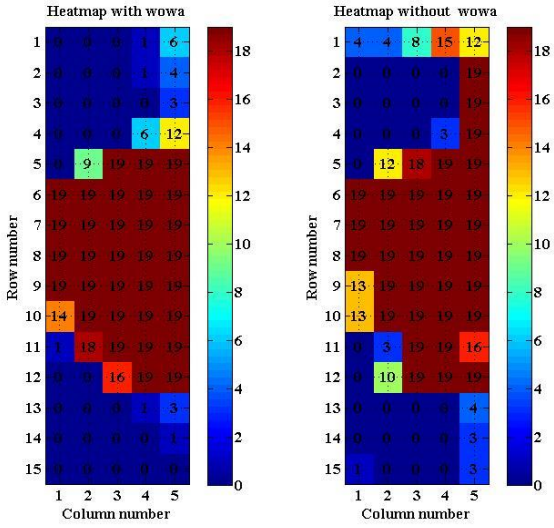


Figure 10
Deformation average test during charge (battery side B)

It can be observed there were a total of 19 deformation states during this test and that the upper and middle parts of the battery changed the most during these measurements. The original data is shown on the left and the results after filtering are shown on the right. It can be seen that the critical points and regions are better concentrated with the filtering. The differences between sides A and B of the same test can be observed in the following in Fig. 11:

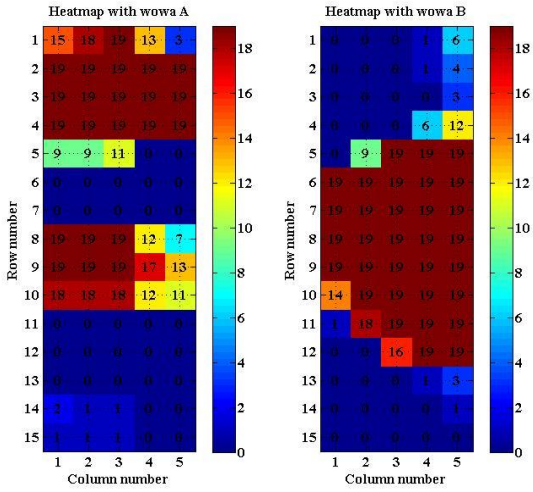


Figure 11
Deformation average test during charge (battery side A and B)

Figure 11 shows the filtered data by using WOWA-based method. We have already established that the cell sides change asymmetrically, and these results support this assumption. In this measurement, the lower part of the battery is the least critical region. In real cases, it is difficult to distinguish exactly which side A and B of the cell are, therefore, the results should be combined. Figure 12 shows the results of merging the two sides during charging and discharging, expressed as a percentage.

Figure 12 shows a charge/discharge cycle for the same battery. The results are similar, but there are also differences. The similarity is that in both cases, 8-9 rows in the middle of the battery, the most critical region, and the bottom of the battery were the least important. The deviation is observed in the upper part of the cell, during discharge, the upper right image became a highly critical area. To better assess the critical areas, all charging and discharging results were projected onto each other. Measurements from all 10 tests (5 charges, 5 discharges) from the three different batteries (sides A and B) were summarized.

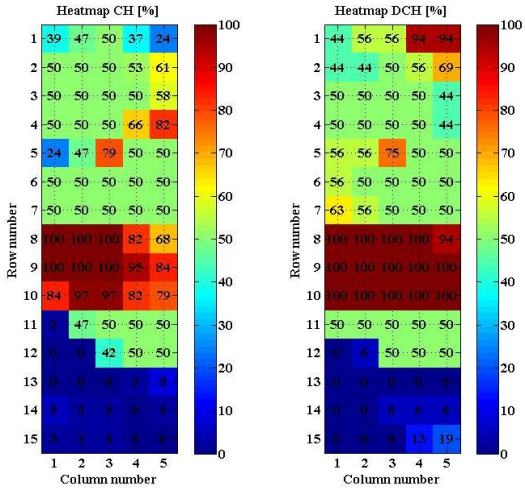


Figure 12
Deformation average test during charge and discharge (combined A and B)

The state after aggregating, averaging and sorting the data is as follows:

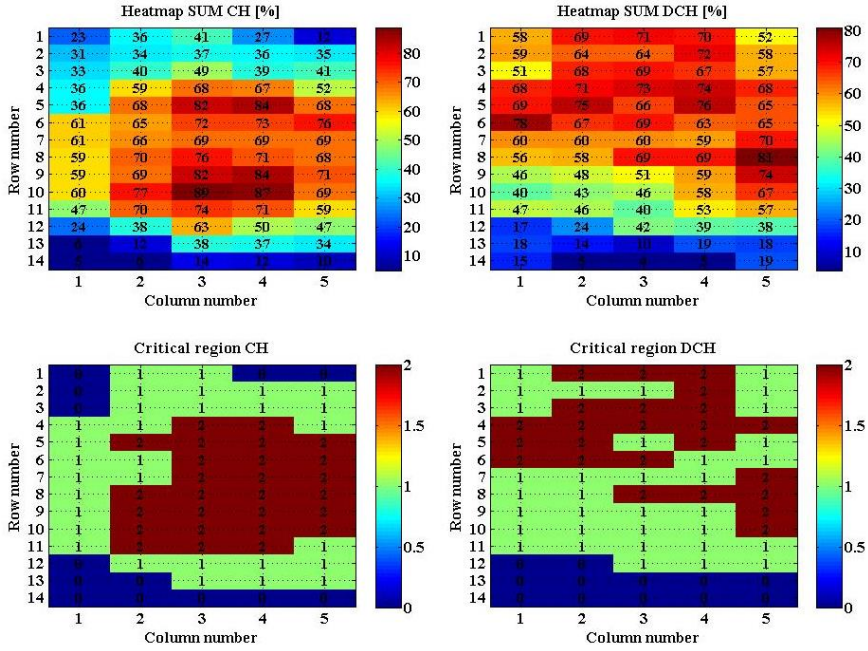


Figure 13
Critical regions of charging and discharging

The combined results of all charge tests for all three cells are shown on the left side of the figure, and the results after discharge are shown on the right side. According to the result after charging, moving towards the middle of the cell, we get the critical areas, the lower and upper parts are less dominant. During discharge, the lower part is also less important, however, in this case, the critical region has shifted to the right and upwards. A summary of these results is required to determine a general case. In terms of cell criticality, the following 3 categories were defined: 0-non-critical (0-33%), 1-moderately critical (33-66%), 2-important region (66-100%). The final step of the study is to combine the results and form a general classification picture. Figure 14 shows the last statement:

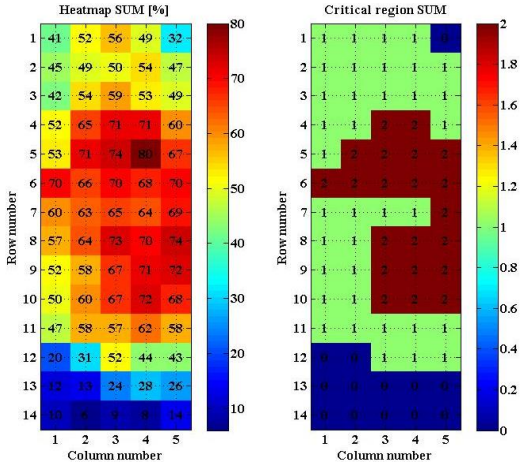


Figure 14
Categorized critical regions of the 5 Ah lithium polymer battery

Based on the images, it can be concluded that the critical regions occur the most common in the middle part of the cell since measurements should concentrate these regions. In the case of measuring in only one or a few points, it is more likely that critical region towards the right edge are found. Furthermore, it can be observed that the center of the cell is rarely the most ideal measuring point. Based on the analysis and calculations, it is advisable to place the displacement sensor above the middle, to the right, 5-6 rows 3-4 columns. The smallest deformation is probably to be measured at the bottom of the cell. In the case of conducting charge (see Fig. 13), test only, the middle part of the battery is the more ideal choice. In case of discharge (see Fig. 13), it is suggested to perform the measurement a little higher than the middle part.

Conclusions

Digitization of the battery surface is a new and state-of-the-art method for obtaining a very accurate data set on the deformation of cells during charging and discharging. Altogether of three lithium polymer battery measurements of the

same type but different states are evaluated. The electrical and surface digitization points are aggregated and arranged in data matrices depending on the charge level. We propose the adoption of WOWA-based filtering method for the detection of critical regions. We effectively filtered the deformation database from errors and outliers, making it easier to identify critical points. Implementing the application of WOWA in a 2D convolution filter to a digital image correlation procedure presents new approach to a current engineering problem. Based on the results, it can be concluded that the method is suitable for filtering deformation data, and critical regions can be efficiently determined on the lithium polymer battery.

Acknowledgment

The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Autonomous Systems National Laboratory Program.

References

[1] D. N. T. How, M. A. Hannan, M. S. Hossain Lipu, and P. J. Ker, "State of Charge Estimation for Lithium-Ion Batteries Using Model-Based and Data-Driven Methods: A Review," IEEE Access, Vol. 7, pp. 136116-136136, 2019

[2] A. Fotouhi, D. J. Auger, K. Propp, S. Longo, and M. Wild, "A review on electric vehicle battery modelling: From Lithium-ion toward Lithium-Sulphur," Renew. Sustain. Energy Rev., Vol. 56, pp. 1008-1021, 2016

[3] Y. Wang et al., "A comprehensive review of battery modeling and state estimation approaches for advanced battery management systems," Renew. Sustain. Energy Rev., Vol. 131, No. July, p. 110015, 2020

[4] F. A. Shah, S. Shahzad Sheikh, U. I. Mir, and S. Owais Athar, "Battery health monitoring for commercialized electric vehicle batteries: Lithium-ion," 5th Int. Conf. Power Gener. Syst. Renew. Energy Technol. PGSRET 2019, no. August, pp. 1-6, 2019

[5] J. Yi, U. S. Kim, C. B. Shin, T. Han, and S. Park, "Three-Dimensional Thermal Modeling of a Lithium-Ion Battery Considering the Combined Effects of the Electrical and Thermal Contact Resistances between Current Collecting Tab and Lead Wire," J. Electrochem. Soc., Vol. 160, No. 3, pp. A437-A443, 2013

[6] Mallarapu, A., Kim, J., Carney, K., DuBois, P., Santhanagopalan, S., Modeling Extreme Deformations in Lithium Ion Batteries, eTransportation

[7] Z. Pan, W. Li, and Y. Xia, "Experiments and 3D detailed modeling for a pouch battery cell under impact loading," J. Energy Storage, Vol. 27, No. August 2019, 2020

[8] J. Zhu, T. Wierzbicki, and W. Li, "A review of safety-focused mechanical modeling of commercial lithium-ion batteries," J. Power Sources, Vol. 378, no. November 2017, pp. 153-168, 2018

[9] L. Greve and C. Fehrenbach, “Mechanical testing and macro-mechanical finite element simulation of the deformation, fracture, and short circuit initiation of cylindrical Lithium ion battery cells,” J. Power Sources, Vol. 214, pp. 377-385, 2012

[10] V. Müller, R. G. Scurtu, M. Memm, M. A. Danzer, and M. Wohlfahrt-Mehrens, “Study of the influence of mechanical pressure on the performance and aging of Lithium-ion battery cells,” J. Power Sources, Vol. 440, no. May, p. 227148, 2019

[11] D. Ren, X. Feng, L. Lu, X. He, and M. Ouyang, “Overcharge behaviors and failure mechanism of lithium-ion batteries under different test conditions,” Appl. Energy, Vol. 250, no. April, pp. 323-332, 2019

[12] C. R. Fell, L. Sun, P. B. Hallac, B. Metz, and B. Sisk, “Investigation of the Gas Generation in Lithium Titanate Anode Based Lithium Ion Batteries,” J. Electrochem. Soc., Vol. 162, No. 9, pp. A1916-A1920, 2015

[13] A. Mukhopadhyay and B. W. Sheldon, “Deformation and stress in electrode materials for Li-ion batteries,” Prog. Mater. Sci., Vol. 63, no. January, pp. 58-116, 2014

[14] W. Liu, H. Liu, Q. Wang, J. Zhang, B. Xia, and G. Min, “Gas swelling behaviour at different stages in Li4Ti5O12/LiNi1/3Co1/3Mn1/3O2 pouch cells,” J. Power Sources, Vol. 369, pp. 103-110, 2017

[15] Y. Li, K. Zhang, B. Zheng, and F. Yang, “Effect of local deformation on the coupling between diffusion and stress in lithium-ion battery,” Int. J. Solids Struct., Vol. 87, pp. 81-89, 2016

[16] K. Kanamura, “Electrolytes for lithium batteries,” Fluorinated Mater. Energy Convers., No. 6, pp. 253-266, 2005, doi: 10.1016/B978-008044472-7/50039-4

[17] Y. Zhan, J. Deng, and T. Wang, “Lithium battery swollen detection based on computer vision,” Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, Vol. 2, pp. 728-731, 2013

[18] B. Bitzer and A. Gruhle, “A new method for detecting lithium plating by measuring the cell thickness,” J. Power Sources, Vol. 262, pp. 297-302, 2014

[19] M. Bauer, M. Wachtler, H. Stöwe, J. V. Persson, and M. A. Danzer, “Understanding the dilation and dilation relaxation behavior of graphite-based lithium-ion cells,” J. Power Sources, Vol. 317, pp. 93-102, 2016

[20] F. Grimsmann, T. Gerbert, F. Brauchle, A. Gruhle, J. Parisi, and M. Knipper, “Determining the maximum charging currents of lithium-ion cells for small charge quantities,” J. Power Sources, Vol. 365, pp. 12-16, 2017

[21] E. Kwak, D. S. Son, S. Jeong, and K. Y. Oh, “Characterization of the mechanical responses of a LiFePO4 battery under different operating conditions,” J. Energy Storage, Vol. 28, no. February, 2020

[22] A. J. Louli, J. Li, S. Trussler, C. R. Fell, and J. R. Dahn, "Volume, Pressure and Thickness Evolution of Li-Ion Pouch Cells with Silicon-Composite Negative Electrodes," J. Electrochem. Soc., Vol. 164, No. 12, pp. A2689-A2696, 2017

[23] B. Rieger, S. V. Erhard, K. Rumpf, and A. Jossen, "A New Method to Model the Thickness Change of a Commercial Pouch Cell during Discharge," J. Electrochem. Soc., Vol. 163, No. 8, pp. A1566-A1575, 2016

[24] K. Y. Oh, B. I. Epureanu, J. B. Siegel, and A. G. Stefanopoulou, "Phenomenological force and swelling models for rechargeable lithium-ion battery cells," J. Power Sources, Vol. 310, pp. 118-129, 2016

[25] T. Cai, S. Pannala, A. G. Stefanopoulou, and J. B. Siegel, "Battery Internal Short Detection Methodology Using Cell Swelling Measurements," Proc. Am. Control Conf., Vol. 2020-July, pp. 1143-1148, 2020

[26] D. Lu, S. Lin, W. Cui, S. Hu, Z. Zhang, and W. Peng, "Swelling mechanism of 0%SOC lithium iron phosphate battery at high temperature storage," J. Energy Storage, Vol. 32, no. July, p. 101791, 2020

[27] J. H. Lee, H. M. Lee, and S. Ahn, "Battery dimensional changes occurring during charge/discharge cycles - Thin rectangular lithium ion and polymer cells," J. Power Sources, Vol. 119-121, pp. 833-837, 2003

[28] B. Sood, C. Hendricks, M. Osterman, and M. Pecht, "Health monitoring of lithium-ion batteries," Electron. Device Fail. Anal., Vol. 16, No. 2, pp. 4-16, 2014

[29] W. Choi, Y. Seo, K. Yoo, T. J. Ko, and J. Choi, "Carbon Nanotube-Based Strain Sensor for Excessive Swelling Detection of Lithium-Ion Battery," 2019 20th Int. Conf. Solid-State Sensors, Actuators Microsystems Eurosensors XXXIII, TRANSDUCERS 2019 EUROSENSORS XXXIII, no. June, pp. 2356-2359, 2019

[30] D. Sauerteig, S. Ivanov, H. Reinshagen, and A. Bund, "Reversible and irreversible dilation of lithium-ion battery electrodes investigated by in-situ dilatometry," J. Power Sources, Vol. 342, pp. 939-946, 2017

[31] T. Polóni, M. A. Figueroa-Santos, J. B. Siegel, and A. G. Stefanopoulou, "Integration of Non-monotonic Cell Swelling Characteristic for State-of-Charge Estimation," Proc. Am. Control Conf., Vol. 2018-June, pp. 2306-2311, 2018

[32] Y. Wu et al., "Analysis of manufacturing-induced defects and structural deformations in lithium-ion batteries using computed tomography," Energies, Vol. 11, No. 4, 2018

[33] V. Yufit, P. Shearing, R. W. Hamilton, P. D. Lee, M. Wu, and N. P. Brandon, "Investigation of lithium-ion polymer battery cell failure using X-ray computed tomography," Electrochem. commun., Vol. 13, No. 6, pp. 608-

610, 2011

[34] H. Popp, M. Koller, M. Jahn, and A. Bergmann, "Mechanical methods for state determination of Lithium-Ion secondary batteries: A review," J. Energy Storage, Vol. 32, no. September, p. 101859, 2020

[35] B. Rieger, S. Schlueter, S. V. Erhard, J. Schmalz, G. Reinhart, and A. Jossen, "Multi-scale investigation of thickness changes in a commercial pouch type lithium-ion battery," J. Energy Storage, Vol. 6, pp. 213-221, 2016

[36] M. Gomercic and D. Winter, "Robot-based 3D imaging in industrial inspection," 2nd IEEE Int. Conf. Ind. Informatics, INDIN'04, pp. 421-424, 2004

[37] J. Peterka, L. Morovič, P. Pokorný, M. Kováč, and F. Hornák, "Optical 3D scanning of Cutting tools," Appl. Mech. Mater., Vol. 421, pp. 663-667, 2013

[38] E. M. C. Jones, M. N. Silberstein, S. R. White, and N. R. Sottos, "In Situ Measurements of Strains in Composite Battery Electrodes during Electrochemical Cycling," Exp. Mech., Vol. 54, No. 6, pp. 971-985, 2014

[39] C. Dai et al., "In situ strain measurements and stress analysis of SiO@C composite electrodes during electrochemical cycling by using digital image correlation," Solid State Ionics, Vol. 331, No. September 2018, pp. 56-65, 2019

[40] P. K. Leung et al., "Real-time displacement and strain mappings of lithium-ion batteries using three-dimensional digital image correlation," J. Power Sources, Vol. 271, pp. 82-86, 2014

[41] J. Luo, C.Y. Dai, Z. Wang, K. Liu, W.G. Mao, D.N. Fang, X. Chen, "In-situ measurements of mechanical and volume change of LiCoO2 lithium-ion batteries during repeated charge–discharge cycling by using digital image correlation", Measurement (2016)

[42] R. Tao, J. Zhu, Y. Zhang, W. L. Song, H. Chen, and D. Fang, "Quantifying the 2D anisotropic displacement and strain fields in graphite-based electrode via in situ scanning electron microscopy and digital image correlation," Extrem. Mech. Lett., Vol. 35, p. 100635, 2020

[43] X. Cheng and M. Pecht, "In situ stress measurement techniques on li-ion battery electrodes: A review," Energies, Vol. 10, No. 5, pp. 1-19, 2017

[44] Y. C. Zhang, O. Briat, J. Y. Deletage, C. Martin, G. Gager, and J. M. Vinassa, "Characterization of external pressure effects on lithium-ion pouch cell," Proc. IEEE Int. Conf. Ind. Technol., Vol. 2018-Febru, pp. 2055-2059, 2018

[45] C. Parthasarathy, S. Thanagasundaram, and T. K. Jet, "Study of applied pressure on open circuit characteristics and capacity of lithium polymer

pouch cells," 2016 Asian Conf. Energy, Power Transp. Electrif. ACEPT 2016, 2017

[46] S. Kocsis Szürke and I. Lakatos, "The lithium polymer battery swelling test with high-precision displacement sensors,"2018 20th International Symposium on Electrical Apparatus and Technologies (SIELA), Bourgas, 2018, pp. 1-4

[47] Y. L. Dong and B. Pan, "A Review of Speckle Pattern Fabrication and Assessment for Digital Image Correlation," Exp. Mech., Vol. 57, No. 8, pp. 1161-1181, 2017

[48] B. Llamazares, "A Behavioral Analysis of WOWA and SUOWA Operators," Int. J. Intell. Syst., Vol. 31, No. 8, pp. 827-851, 2016

[49] B. Llamazares, "An analysis of some functions that generalizes weighted means and OWA operators," Int. J. Intell. Syst., Vol. 28, No. 4, pp. 380-393, 2013

[50] L. A. Zadeh, Studies in Fuzziness and Soft Computing: Foreword, Vol. 261. 2010

[51] E. Damiani, S. De Capitani di Vimercati, P. Samarati, and M. Viviani, "A WOWA-based Aggregation Technique on Trust Values Connected to Metadata," Electron. Notes Theor. Comput. Sci., Vol. 157, No. 3, pp. 131-142, 2006

[52] L. Torres, J. C. Becceneri, C. C. Freitas, S. J. S. S. Anna, and S. Sandri, "WOWA image filters."

[53] B. Bede and I. Rudas, "Shooting method for fuzzy two-point boundary value problems," Fuzzy Information Processing Society (NAFIPS), 2012 Annual Meeting of the North American, 2012.08.06-2012.08.08. Berkeley: North American Fuzzy Information Processing Society, pp. 5-8, 2012

[54] C. Carlsson and R. Fuller, Fuzzy Reasoning in Decision Making and Optitnization. Springer-Verlag Berlin Heidelberg, ISBN 978-3-7908-2497-1, 2002

[55] L. Zadeh, "Fuzzy sets," Information and Control, Vol. 8, pp. 338-353, 1965

[56] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," IEEE Transactions on System, Man and Cybernetics, Vol. 18, pp. 183-190, 1988

[57] G. Beliakov, A. Pradera, and T. Calvo, Aggregation Functions: A Guide for Practitioners. Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-73720-9, 2007

[58] G. Choquet, "Theory of capacities," Annales de l'Institut Fourier, Vol. 5, pp. 131-154, 1953

[59] V. Torra, "The weighted owa operator," International Journal of Intelligent Systems, Vol. 12, pp. 153-166, 1997

Complex Expert Assessment of the State of Business Enterprises

Romualdas Ginevicius[1], Dainora Gedvilaite[2], Andrius Stasiukynas[3], Karel Suhajda[4]

[1]Vilnius Gediminas Technical University, Institute of Dynamic Management, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania, romualdas.ginevicius@vilniustech.lt

[2]Vilniaus Kolegija/University of Applied Sciences, Faculty of Economics, Saltoniškių g. 58, LT-08105 Vilnius, Lithuania, d.gedvilaite@ekf.viko.lt

[3]Kazimieras Simonavičius University, Dariaus ir Girėno g. 21 LT-02189 Vilnius, Lithuania, andrius.stasiukynas@ksu.lt

[4]Brno University of Technology, Antonínská 548/1, 60190 Brno, Czech Republic, suhajda.k@fce.vutbr.cz

Abstract: The commercial success of businesses depends to a large extent on the ability to quantify the current situation at the desired point in time. It helps to make the right strategic development decisions and reveals weaknesses. The state of development of a company is a complex phenomenon, therefore, it can be described only by a certain number of indicators. They are multidimensional and of unequal importance, therefore, multiple-criteria methods are used to combine their values into one generalising quantity. They rely to a large extent on expert assessments to determine the weights of indicators, as well as the values of indicators that are difficult to formalise. An integral part of such assessments is the examination of the consistency of expert opinions. Existing methods for determining the level of consistency of expert assessment are intended to determine the importance of indicator weights, and it is not possible to determine the consistency of expert assessment of indicator values on their basis. This is because, when determining the importance of an indicator, the estimate of the importance of one indicator follows from the context of the importance of all other indicators, whereas in the second case, the value of each indicator is determined separately, i.e., it does not follow from the context of the values of the other indicators. In order to determine the consistency of the expert assessment of the values of indicators, it is necessary to calculate the actual and maximum possible level of uniformity or non-uniformity of the assessment. The consistency of the expert assessment will be demonstrated by the ratio of these values. The aim of the article is to propose and approve a methodology for determining the compatibility of expert assessment of the values of difficult-to-formalize indicators that increase the commercial success and competitiveness of business enterprises.

Keywords: business enterprises; multiple-criteria methods; compatibility of expert assessments

1 Introduction

In the context of global economic integration, economic operators seek to be on an equal footing in international markets. This is especially true for Eastern and partly for Central European countries. The main challenge for their businesses is to be competitive. Competitiveness is often understood as the share in both foreign and domestic markets. It is an integral result of business development. Assessing the importance of competitiveness is comprehensively examined in many scientific studies. Basically, they are all conducted for one purpose – finding the ability to change the state of a business, i.e., to pursue its success. In other words, ways to manage business development in a targeted manner are being sought. Global scientists Drucker and Sukhart have identified an essential condition for solving this problem – the development process can be managed if there is a possibility to quantify its condition at a desired time. Thus, in order to improve business results, at least two things are required: first, to know what business development depends on; and second, how to quantify its current situation. Only then can the right business development strategic decisions be made.

Solving these tasks is not easy. This is because business development is a complicated complex socio-economic process, which includes both the complex of interacting people and the necessary material and technical and other resources – materials, equipment, technologies, information flows, etc. This allows business to be seen as a socio-economic system. Due to their complexity, such systems manifest themselves in reality in a number of aspects of the most diverse nature. When formalised, they become indicators that reflect the state of the system. They can be expressed in different dimensions and vary in opposite directions. This means that an increase in some of their values improves the situation, while increases in others worsens it. For example, the higher a company's advertising costs, the higher the expected sales of products or services and the better the performance. Rising advertising costs of competitors can make these results worse. In addition to the above, another factor is no less important – some indicators can be easily formalised, while others – with a degree of difficulty. The said advertising costs can be accurately estimated by the amount of money spent for that purpose. Meanwhile, it is not possible to accurately "measure", for example, a company's ability to assess the competitors' market behaviour. Another problematic aspect to consider when assessing a business situation is the unequal importance of the factors that affect it.

In order to get a general picture of the business situation, it is necessary to combine important indicators expressing the contradictory factors that affect it into one generalising index. Research in recent years has shown that multiple-criteria approaches are best suited to address this issue. Long-term practice has highlighted two conditions for their application: first, indicators expressed in different dimensions and moving in the opposite direction need to be made comparable with each other, i.e., dimensionless, and their relevance to the phenomenon in question, i.e., business development, must be quantified. Knowing the normalised values of

the indicators and their importance, the simplest way to obtain the generalised index is to sum up the sum the product of the values and importance of the indicators, or by other developed multiple-criteria assessment methods.

The weights of the indicators, as well as the values of the indicators that are difficult to formalise, are determined by experts. It is no coincidence that in the theory of multiple-criteria assessment, they are given a special role. The purpose of the expert assessment is to reflect the situation under assessment as adequately as possible. On the other hand, despite the professionalism of the team of experts, their opinions tend to differ. In order to use the obtained results in further calculations, it is necessary to determine whether they are consistent. A number of studies and various methodologies have been devoted to this problem. Their analysis shows that they all address the issues of consistency in the expert assessment of indicator weights. Meanwhile, the evaluation of indicator values have practically not been performed. The average of the estimates is used for further calculations. This may be due to the fact that no appropriate methodologies have been proposed. In this situation, the question arises as to whether the methods of consistency of expert assessment of indicator weights are also suitable for determining the consistency of assessment of indicator values. If not, by increasing the adequacy of the application of multiple-criteria methods to the assessment of the state of business enterprises, appropriate ways to determine consistency need to be sought.

2 Literature Review

The main purpose of a business is profit making, except for non-profit organisations. To achieve it, a group of stakeholders forms a structure, i.e., creates production and organisational/management staff, and a social system, into which it integrates the necessary material, technical and other resources. This structure, once validated, acquires the appropriate status of an official organization or company. Thus, structure can be seen as a means to an end. It can only be achieved through the constant development of the company [1]. The targeted development process does not happen by itself – it needs to be managed. This requires being able to quantify its condition at a desired point in time.

The problem of quantifying the condition of socio-economic systems, such as business enterprises, has received a lot of attention in research. By their nature, they are large and complex, thus, the number of indicators that reflect their situation can be quite large. Combining them into one generalising quantity or index is difficult due to their contradictory nature – they are of different dimensions, can change in opposite directions, some of them must reflect difficult to formalise factors and are not equally important for the phenomenon under consideration. In recent years, multiple-criteria methods have been successfully applied to calculate such indices. The most well known and most widely used are SAW [2], TOPSIS [3], VIKOR [4],

COPRAS [5, 6], ELECTRE III, ELECTRE IV [7, 8], PROMETHEE [9], LINMAP [10], MOORA, MULTIMOORA ([11, 39] and others.

The essence of multiple-criteria assessments is most clearly reflected by the SAW method [3]:

$$K_j = \sum_{i=1}^{n} w_i \widetilde{q_{ij}} \quad (1)$$

where K_j – significance of the multiple-criteria evaluation of the j[th] alternative of the analysed phenomenon by the SAW method; w_i – weight of the i[th] indicator; $\widetilde{q_{ij}}$ – normalised value of j[th] alternative of i[th] indicator; n – number of indicators ($i=\overline{1,n}$).

As can be seen from formula (1), in multiple-criteria evaluations, each indicator is expressed in two quantities: importance and significance [3, 12, 13, 14, 15, 16].

It is true that there are also methods of multiple-criteria evaluation, where indicators are expressed only in terms of significance, i.e., without weights [17, 18, 19, 20, 21].

Indicator weights are usually determined on the basis of subjective methods, i.e., their importance is determined by experts [3, 22, 23, 19, 24, 25, 26, 27, 28, 29]. The objective weight determination methods are also sometimes used, where the importance of an indicator is determined by its significance [30, 31].

The situation is different with the significances of indicators. They can be conditionally divided into two groups. The first group are easily formalised indicators. Their significances are determined precisely because all the necessary information can be found in various sources – statistical publications, normative materials, different reports, project documentation, or simply determined through calculations. Indicators that are difficult to formalise include those whose significance in the chosen evaluation system can be determined only by experts. All this variety of setting indicator weights and significances is demonstrated in Figure 1.

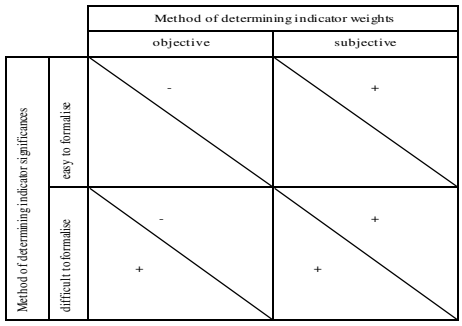


Figure 1

The need for expert assessments in determining the weights and significances of multiple-criteria assessment indicators of the state of SES development (source: compiled by the authors)

As demonstrated in Figure 1, it appears that in only one of the four possible cases it is not necessary to use expert assessments at all. Thus, their importance in multiple-criteria assessments is high.

Determining the level of consistency of expert opinions is inseparable from expert assessments. This important issue is given a lot of attention in the theory of expert assessment, because the results of assessment can be used in further calculations only if the opinions of experts are consistent.

The analysis of literature sources shows that research to determine the consistency of opinions is predominant [32, 33, 34, 35, 36, 37]. It makes sense to see whether the proposed methodologies are also suitable for determining the consistency of expert assessment of indicator values.

In recent years, the methodologies for assessing the consistency of expert assessment proposed by Kendall and Saaty have been the most widely used. The methodology proposed by Saaty is specific, it is integrated into the AHP method for determining indicator weights, therefore, it cannot be treated as an independent or universal methodology [38]. The commonly used Kendall method has a different nature. In this case, the expert assessment of the importance of the indicators is based on the matrix $R=\|q_{ij}\|$. It lists the assessments of importance of each indicator q_{ij} (q_{ij} – is the score of the i^{th} indicator given by the j^{th} expert in the adopted evaluation system, $i = \overline{1,m}$, $j = \overline{1,r}$, where m – is the number of indicators, r – is the number of experts) [32] (Figure 2).

$$Q = \begin{vmatrix} q_{11} & q_{12} & q_{13} & \cdots & q_{1j} & \cdots & q_{1r} \\ q_{21} & q_{22} & q_{23} & \cdots & q_{2j} & \cdots & q_{2r} \\ q_{31} & q_{32} & q_{33} & \cdots & q_{3j} & \cdots & q_{3r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{i1} & q_{i2} & q_{i3} & \cdots & q_{ij} & \cdots & q_{ir} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{m1} & q_{m2} & q_{m3} & \cdots & q_{mj} & \cdots & q_{mr} \end{vmatrix}$$

Figure 2
Matrix of expert assessment of the importance of indicators

The next step in the consistency assessment is ranking of Q estimates in the matrix (Fig. 3). The result is matrix R, which shows the ranks of all indicators (r_{ij} – the rank of ith indicator given by the jth expert).

$$R = \begin{vmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1j} & \dots & r_{1r} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2j} & \dots & r_{2r} \\ r_{31} & r_{32} & r_{33} & \dots & r_{3j} & \dots & r_{3r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{i1} & r_{i2} & r_{i3} & \dots & r_{ij} & \dots & r_{ir} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & r_{mj} & \dots & r_{mr} \end{vmatrix}$$

Figure 3
Matrix of expert assessment of the importance of indicators

The method for determining the consistency of expert assessments depends on the number of indicators. If it does not exceed 7 ($m \leq 7$), the degree of compatibility is indicated by the coefficient W:

$$W = \frac{S_f}{S_{max}} \tag{2}$$

$$S_f = \sum_{i=1}^{m} \left(\sum_{j=1}^{r} r_{ij} - \bar{r} \right)^2 \tag{3}$$

$$\bar{r} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{r} r_{ij}}{m} \tag{4}$$

$$S_{max} = \frac{r^2 m(m^2-1)}{12} \tag{5}$$

where S_f – the sum of the deviations of r_i values from the mean r_{ij}; $\bar{r}$ – the total average of r_{ij} values; S_{max} – the maximum possible deviation of r_i values from the mean r_{ij}.

The closer the value of the coefficient W is to 1, the more consistent the expert opinion.

When the number of evaluated indicators is greater than 7 ($m > 7$), the Pearson correlation factor χ^2 is calculated:

$$\chi_f^2 = \frac{12 S_f}{rm(m+1)} \tag{6}$$

The actual value of χ_f^2 coefficient obtained is compared with the critical value χ_{kr}^2, which depends on the number of degrees of freedom $\gamma (\gamma = m - 1)$ and the α level of significance chosen (in the social science, the α value is usually equal to 0.05, i.e., (1-0.05)x100 = 95 %). If $\chi_f^2 > \chi_{kr}^2$, the opinions of the experts are consistent.

The analysis of Kendall's method for determining the consistency of expert assessment allows to draw an essential conclusion – the importance estimate of i^{th} indicator follows from all other estimates of assessment of this indicator. In other words, when giving an estimate of importance of i^{th} indicator, the expert must weigh its relation to the importance of all other indicators for the phenomenon in question.

This is very clearly demonstrated by the essential condition for determining the importance of indicators, which looks as follows [3]:

$$\omega_1 + \omega_2 + \omega_3 + \ldots + \omega_i + \ldots + \omega_m = \sum_{i=1}^{\min} \omega_i = 1.0 \quad (7)$$

Therefore, in determining the weights of indicators of the phenomenon under consideration, the rows of the matrix in Fig. 2 are interrelated. It is from this condition that all known subjective methods for determining indicator weights arise [32, 22, 23, 19, 24, 26, 30].

The situation is different with the determination of consistency of expert assessment of the values of indicators that are difficult to formalise. Unlike determining the consistency of the expert assessment of indicator weights, in this case the value of the indicator in question does not depend on the estimates of other indicators. In this case, the expert assessment of i^{th} indicator will be based on a matrix $\mathbf{Q}_i$ with only one row (Fig. 4).

$$\mathbf{Q}_i = |p_{i1} \quad p_{i2} \quad p_{i3} \quad \ldots \quad p_{ij} \quad \ldots \quad p_{ir}|$$

Figure 4

Matrix of expert evaluation of the value of ith indicator of the analysed phenomenon

Figures 2 and 3 demonstrate that the methods for determining the consistency of expert assessment of the importance of indicators are not suitable for determining the consistency of expert assessment of indicator values.

3 Methodology

The desired level of compatibility of the expert assessment of the indicator values must meet the following requirements:

1) The value of the desired indicator must vary from 0 to 1.

2) If the opinion of all experts is unanimous, i.e., if they all give the same estimate for the indicator in question, the value of the compatibility indicator must be equal to 1.

3) If the experts gave the most divergent estimates on the adopted rating scale to the indicator in question, the value of the compatibility indicator must be close to 0.

4) It must be possible to quantify both the actual and the maximum possible uniformity of expert opinions depending on the specific situation, i.e., both from the indicators of the phenomenon in question and the number of experts evaluating them.

5) When assessing the limitations of points 1 to 4, the required indicator must be determined as the ratio of the actual and the maximum possible uniformity of the estimates provided by the experts to the specific situation.

The proposed indicator of the level of consistency of the expert assessment of the values of indicators must meet all the applicable requirements:

$$W_i = 1 - \frac{W_i^f}{W_i^{max}} \text{ or } W_i = \frac{W^{max} - W_i^f}{W^{max}} \quad (8)$$

where W_i – the indicator of consistency of the expert assessment of the value of the indicator in question; W_i^f – the actual indicator of non-uniformity in the expert opinions of indicator i[th]; W_i^{max} – the most likely indicator of non-uniformity of expert opinions of indicator i[th].

As the formula (7) shows, in the ideal situation, i.e., when $W_i^f = W_i^{max}$, $W_i = 0$, as the W_i^f value approaches 0, the value of the compatibility indicator W_i also approaches 1. Consequently, it ranges from 0 to 1.0.

Thus, the task of determining the degree of compatibility of an expert assessment consists of determination of values W^f and W^{max}.

Determination of actual non-uniformity, i.e., value W_i^f „ of the expert assessments of i[th] indicator of the analysed phenomenon. The degree of non-uniformity of expert assessment depends on the degree of dispersion of the expert assessment estimate, therefore, W^f will be calculated as follows:

$$W_i^f = \sum_{i=1}^{r} \left| \bar{q}_i - q_{ij} \right| \quad (9)$$

where $\bar{q}_i$ – the average of the expert assessment of the i[th] indicator value.

The value $\bar{q}_i$ is determined as follows:

$$\bar{q}_i = \frac{\sum_{i=1}^{r} q_{ij}}{r} \quad (10)$$

Formula (9) demonstrates that if the estimates of all the experts are the same, then $W_i^f = 0$.

Determination of maximum non-uniformity, i.e., value W_i^{max}, of the expert assessments of i[th] indicator of the analysed phenomenon.

The value W_i^{max} reflects the situation where the expert assessment estimates differ the most. In this case, the matrix $\tilde{q}_i = \left\| q_{ij} \right\|$ of expert assessment of significance of i[th] indicator of the analysed phenomenon will look as follows (Fig. 5).

$$\tilde{q}_i = \left|\tilde{q}_{i1}^{max}\, \tilde{q}_{i2}^{min}\, \tilde{q}_{i3}^{max}\, \tilde{q}_{i4}^{min} \ldots \tilde{q}_{ij-1}^{max}\, \tilde{q}_{ij}^{min}\, \tilde{q}_{ij+1}^{max} \ldots \tilde{q}_{ir-1}^{max}\, \tilde{q}_{ir}^{min}\right|$$

Figure 5

Matrix of expert assessment of the significance of ith indicator of the phenomenon in question, when the opinions of the experts differ the most (q_ij^max – the highest possible estimate of significance of i[th] indicator given by the jth expert; q_ij^min – the same, the lowest possible estimate)

As with the determination of W^f, W^{max} shall be calculated as follows:

$$W_i^{max} = \sum_{i=1}^{r}\left|\tilde{\tilde{q}}_i - \tilde{q}_{ij}\right| \tag{11}$$

The value $\tilde{\tilde{q}}_i$ is determined as follows:

$$\tilde{\tilde{q}}_i = \frac{\sum_{i=1}^{r}\tilde{q}_i}{r} \tag{12}$$

where $\tilde{\tilde{q}}_i$ – the average of the expert assessment of the value of i[th] indicator, when the expert opinions differ the most; $\tilde{q}_i$ – the significance of i[th] indicator, when expert opinions differ the most.

It is not difficult to notice that when the number of experts is even, then W_i^{max}= $(\tilde{\tilde{q}}_i - 1) \times r$. Table (1) presents the values of W^{max} when r varies from 2 to 10.

Table 1

Values of Wmax significance depending on the number of experts

Number of experts	2	3	4	5	6	7	8	9	10
Value W^{max} significance	9.00	12.00	18.00	21.00	27.00	30.86	36.00	40.00	45.00

The value W^{max}, like the value W^f, essentially reflects the greatest possible degree of non-uniformity depending on the number of expert's r.

4 Empirical Research

The proposed methodology for determining the consistency of expert assessment of the indicator significance has been tested on the basis of a real example. The experts provided the following estimates of difficult-to-formalize indicators of the commercial success of the examined business in a ten-point system (Table 2).

The three indicators (age of company, advertising costs and new product development costs) in Table 2 are easy to formalise because their values are known precisely, while all other indicators are difficult to formalise, the values of which have been determined by experts. It is necessary to assess whether their opinions are consistent based on the proposed methodology.

Table 2

Estimates of the evaluated company's commercial success indicators and the results of calculation of consistency of the expert assessments

Measurement unit	Row No	Indicator	$\bar{q}_i$	W^f	W^{max}	W_i	Consistency assessment results
Score points	1	Country's level of infrastructure	5.75	34	36	0.61	Inconsistent
Years	2	Age of company	-	-	-	-	-
Score points	3	Appropriate business development strategy	7.75	7.0	36	0.82	Consistent
Euro	4	Advertising costs	-	-	-	-	-
Score points	5	Employee incentive system	5.75	8.0	36	0.78	Consistent
Score points	6	Quality of products	7.25	10	36	0.72	Inconsistent
Score points	7	Packaging of products	4625	9.0	36	0.75	Consistent
Score points	8	Country's level of economic development	6.00	11	36	0.69	Inconsistent
Euro	9	New product development costs	-	-	-	-	-
Score points	10	Manager competence	5.00	8.0	36	0.78	Consistent

Determining the value W_i^f of the actual non-uniformity of expert assessment of the indicator significance.

According to formula (9), the average of the expert assessment of the values of the first indicator, the country's level of infrastructure, will be equal to:

$$\bar{q}_1 = \sum_{i=1}^{r} \frac{q_{ij}}{r} = \frac{46}{8} = 5.75$$

The value W_1^f according to formula (9) will be equal to:

$$W_1^f = (5.75 - 4) + (5.75 - 4) + (5.75 - 8) + (5.75 - 3) + (5.75 - 7) + (5.75 - 6) + (5.75 - 5) + (5.75 - 9) = 14.$$

Determining the value W_i^{max}, of the maximum non-uniformity of expert assessment of the indicator significance.

The situation of expert assessment of the indicator significance, when expert opinions are most inconsistent, will look as follows (Table 3):

Table 3

Maximum and minimum values of expert assessment (scale 10 points)

Experts	1	2	3	4	5	6	7	8	Total
Estimates	10	1	10	1	10	1	10	1	44

The average of the expert assessment of the significance of the first indicator, based on formula (12) and Table 3, will be equal to:

$$\tilde{\tilde{q}}_i = \frac{44}{8} = 5.5$$

The value W_1^{max}, according to formula (11) will be equal to:

$$W_1^{max} = \sum_{i=1}^{r} \lfloor \tilde{\tilde{q}}_1 - \tilde{q}_1 \rfloor = 4 \times |5.5 - 10| + 4 \times |5.5 - 1| = 36, arba\ W_1^{max} = (5.5 - 1) \times 8 = 36$$

Mean significance $\tilde{\tilde{q}}_i$ and value W_i^{max} significance were determined in the same way (Table 2).

According to formula (8), the value W_i for the first indicator will be equal to:

$W_1 = 1 - \frac{34}{36} = 0.06$; $W_1 = \frac{36-34}{36} = 0.06$.

The significances of value W_i are presented in Table 2. It demonstrates that four of the seven indicators assessed are consistent, while three are not. Another conclusion is that the opinions of the experts differed the most in those indicators that reflect the external environment of the business enterprise. The critical limit of the consistency coefficient W is 0.75. It is based on empirical research and expert evaluations. The performed calculations confirmed the suitability of the proposed methodology for expert assessment of significances of difficult-to-formalise indicators.

5 Discussion

In multiple-criteria evaluations, the expert assessments of significance of the difficult-to-formalise indicators have some unresolved issues. Meanwhile, there may be a number of indicators that reflect the development aspects of a business enterprise. This is illustrated by the example presented in this article. The analysis

of the literature sources dedicated to this question, including the defended dissertations, demonstrated that such an assessment is hardly ever performed. This issue is ignored. The simplest way taken – determining the arithmetic mean of the expert estimates and treating it as the significance of the indicator sought. This has a strong impact on the adequacy of multiple-criteria assessment.

Undoubtedly, there may be other methods for determining the consistency of expert assessment of significance of difficult-to-formalise indicators. The mathematical statistics methods could open up wide possibilities for this, but their application is limited by insufficient statistical information. This is due to the fact that the system of indicators of the phenomenon in question usually consists of too few indicators in terms of mathematical statistical methods. If their number is large enough, the related indicators are grouped to increase the adequacy of the assessment. This allows forming a hierarchical system of indicators, which makes it possible to reduce the number of indicators evaluated simultaneously.

It can be expected that when this problem is understood to the required extent, it will receive more attention and more reasoned suggestions will be offered.

The proposed methodology can find wide application both in research and in practice, as business people today are increasingly beginning to understand the importance of strategic planning. Quantitative assessment of the current state of the business plays an important role in this process. It is necessary for forecasting changes, solving problems of sustainability of enterprise development, etc.

Conclusions

The experts have an important role to play in multiple-criteria quantitative assessments of the state of development of socio-economic systems, such as businesses. They help to determine the weights of indicators of the phenomenon in question, as well as the significance of the indicators that are difficult-to-formalise. An integral part of such assessments is the determination of the level of consistency of expert opinions. The most common and widely used are the methods for determining the level of consistency of expert assessment of the importance of indicators, while less attention is paid to the consistency of expert assessment of indicator significance. Meanwhile, the methods of expert assessment of the importance of indicators are not suitable for determining the level of consistency of expert assessment of indicator significances. This is because in the case of determining the level of consistency of the expert assessment of indicator weights, the assessment of the importance of the indicator in question is derived from all other estimates of the assessment of the importance of this indicator. Whereas, in the case of expert assessment of the indicator significances, the significance of the analysed indicator does not depend on the estimates of other indicators.

The proposed indicator of the level of consistency of the expert assessment of the significance of the indicator in question ranges from 0 to 1, and thus, reflects the extremes of the assessment, when the expert opinions are completely uniform or

completely different. The indicator is defined as the ratio of the actual and the maximum possible uniformity of the estimates provided by the experts to the specific situation, which is characterised by the number of indicators and experts. The performed calculations confirmed the appropriateness of the proposed methodology. It is universal and can be used in quantitative assessments of the state of development of a wide variety of phenomena.

References

[1] S. G. Anton, M. Carp, M: The effect of discretionary accruals on firm growth. Empirical evidence for SMEs from emerging Europe. Journal of Business Economics and Management, Vol. 21, 2020, No. 4, pp. 1128-1148, https://doi.org/10.3846/jbem.2020.12734

[2] K. R. MacCrimmon: Decisionmaking Among Multiple-Attribute Alternatives. A Survey and Consolidated Approach. RAND Memorandum RM-4823-ARPA, 1968

[3] C.-L. Hwang, K. Yoon: Multiple Attribute Decision Making. Methods and Applications a State-of-the-Art Survey. Lecture Notes in Economics and Mathematical Systems 186, Springer Berlin Heidelberg, 1981

[4] S. Opricovic, G.-H. Tzeng: (). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. European Journal of Operational Research, Vol. 156, 2004, No. 2, pp. 445-455, https://doi.org/10.1016/S0377-2217(03)00020-1

[5] E. K. Zavadskas, A. Kaklauskas, N. Banaitienė: Multivariant design and multiple criteria analysis of a building life cycle. Informatica. Vol. 12, 2001, No. 1 (2001) pp. 169-187

[6] A. Kaklauskas, E. K. Zavadskas, S. Raslanas: Multivariant design and multiple criteria analysis of building refurbishments. Energy and Buildings, Vol. 37, 2005, No. 4, pp. 361-372

[7] B. Roy: Revisiting carrying capacity: Area-based indicators of sustainability (la methode Electre). Revue Francaise d'informatique et de Recherrche Operationnelle, Vol. 8, 1968, pp. 57-75

[8] B. Roy: Multicriteria methodology for decision aiding. Dordrecht, Netherlands; Boston, Mass.: Kluwer Academic Publishers. 1996

[9] J. P. Brans, B. Mareschal, P. Vincke: "PROMETHEE: A new family of outranking methods in multicriteria analysis". Operational Research, Vol. 3, 1984, pp. 477-490

[10] V. Srinivasan, A. D. Shocker: Linear programming techniques for multidimensional analysis of preferences. Psychometrika, Vol. 38, 1973, pp. 1973337-369, https://doi.org/10.1007/BF02291658

[11] W. K. M. Brauers, R. Ginevičius: The economy of the Belgian regions tested with MULTIMOORA. Journal of Business Economics and Management, Vol. 11, 2010, No. 2, pp. 173-209

[12] C. W. Churchman, R. L. Ackoff, N. M Smith: an Approximate Measure of Value. Journal of the Operations Research Society of America, Vol. 2, 1954, No. 2, pp. 172-187

[13] A. J. Klee: The Role of Decision Models in the Evaluation of Competing Environmental Health Alternatives Competing Environmental Health Alternatives. Management Science, Vol. 18, 1971, No. 2, pp. 52-67, https://www.jstor.org/stable/2629528

[14] K. R. MacCrimmon: Decisionmaking Among Multiple-Attribute Alternatives. A Survey and Consolidated Approach. RAND Memorandum RM-4823-ARPA, 1968

[15] V. Srinivasan, A. D. Shocker: Linear programming techniques for multidimensional analysis of preferences. Psychometrika, Vol. 38, 1973, pp. 337-369, https://doi.org/10.1007/BF02291658

[16] M. Lisiński, A. Augustinaitis, L. Nazarko, S. Ratajczak: Evaluation of dynamics of economic development in Polish and Lithuanian regions. Journal of Business Economics and Management, Vol. 21, 2020, No. 4, pp. 1093-1110, https://doi.org/10.3846/jbem.2020.12671

[17] A. Arabsheybani, M. M. Paydar, A. S. Safaei: An integrated fuzzy MOORA method and FMEA technique for sustainable supplier selection considering quantity discounts and supplier's risk. Journal of Cleaner Production, Vol. 190, 2018, pp. 577-591, https://doi.org/10.1016/j.jclepro.2018.04.167

[18] R. Dabbagh, S. Yousefi: A hybrid decision-making approach based on FCM and MOORA for occupational health and safety risk analysis. Journal of Safety Research, Vol. 71, 2019, pp. 111-123, https://doi.org/10.1016/j.jsr.2019.09.021

[19] H. Dinçer, S. Hüksel, L. Martínez: Interval type 2-based hybrid fuzzy evaluation of financial services in E7 economies with DEMATEL-ANP and MOORA methods. Applied Soft Computing, Vol. 79, 2019, pp. 186-202, https://doi.org/10.1016/j.asoc.2019.03.018

[20] R. K. Mavi, M. Goh, N. Zarbakhshnia: Sustainable third-party reverse logistic provider selection with fuzzy SWARA and fuzzy MOORA in plastic industry. International Journal of Advanced Manufacturing Technology, Vol. 91, 2017, pp. 2401-2418, https://doi.org/10.1007/s00170-016-9880-x

[21] S. Mete: Assessing occupational risks in pipeline construction using FMEA-based AHP-MOORA integrated approach under Pythagorean fuzzy environment. Human and Ecological Risk Assessment: An International Journal Vol. 25, 2019, No. 7, pp. 1645-1660, https://doi.org/10.1080/10807039.2018.1546115

[22] S. Yousefzadeh, K. Yaghmaeian, A. Hossein Mahvi, S. Nasseri, N. Alavi, R. Nabizadeh: Comparative analysis of hydrometallurgical methods for the recovery of Cu from circuit boards: Optimization using response surface and selection of the best technique by two-step fuzzy AHP-TOPSIS method. Journal of Cleaner Production, Vol. 249, 2020, pp. 119-401, https://doi.org/10.1016/j.jclepro.2019.119401

[23] L. A. Ocampo, C. M. Himang, A. Kumar: a novel multiple criteria decision-making approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy AHP for mapping collection and distribution centers in reverse logistics. Advances in Production Engineering & Management, Vol. 14, 2019, No. 3, pp. 297-322, https://doi.org/10.14743/apem2019.3.329

[24] S. Boral, S. I. Howard, S. K. Chaturvedi, K. McKee, V. N. A. Naikan: An integrated approach for fuzzy failure modes and effects analysis using fuzzy AHP and fuzzy MAIRCA. Engineering Failure Analysis, Vol. 108, 2020, pp. 104-195

[25] J. Seyedmohammadi, F. Sarmadian, A. A. Jafarzadeh, R. W. McDowell: Integration of ANP and Fuzzy set techniques for land suitability assessment based on remote sensing and GIS for irrigated maize cultivation. Archives of Agronomy and Soil Science, Vol. 65, 2018, No. 8, pp. 1063-1079, https://doi.org/10.1080/03650340.2018.1549363

[26] Z. Chen, X. Ming, X. Zhang, D. Yin, Z. Sun: A rough-fuzzy DEMATEL-ANP method for evaluating sustainable value requirement of product service system. Journal of Cleaner Production, Vol. 228, 2019, pp. 485-508, https://doi.org/10.1016/j.jclepro.2019.04.145

[27] A. Bathaei, A. Mardani, T. Balezentis, S. R. Awang, D. Streimikiene, G. C. Fei, N. Zakuan: Application of Fuzzy Analytical Network Process (ANP) and VIKOR for the Assessment of Green Agility Critical Success Factors in Dairy Companies. Symetry, Vol. 11, 2019, p. 2520, doi: 10.3390/sym11020250

[28] Z. Ayağ, F. Samanlioglu: Fuzzy AHP-GRA approach to evaluating energy sources: a case of Turkey. International Journal of Energy Sector Management, Vol. 14, 2018, No. 1, pp. 40-58, https://doi.org/10.1108/IJESM-09-2018-0012

[29] A. R. Pilevar, H. R. Matinfar, A. Sohrabi, F. Sarmadian: Integrated fuzzy, AHP and GIS techniques for land suitability assessment in semi-arid regions for wheat and maize farming. Ecological Indicators, Vol. 110, 2020, 105887, https://doi.org/10.1016/j.ecolind.2019.105887

[30] V. Podvezko: The Comparative Analysis of MCDA Methods SAW and COPRAS. Engineering Economics, Vol. 22, 2011, No. 2, pp. 134-146, http://dx.doi.org/10.5755/j01.ee.22.2.310

[31] V. Podvezko, A. Podviezko: Methods of estimation of weights. Lietuvos matematikos rinkinys [Lithuanian Mathematics Collection], Vol. 55, 2014, pp. 111-116

[32] M. G. Kendall: Rank Correlation Methods. 4th edition. London: Charles Griffin, 1975

[33] T. L. Saaty: The Analytic Hierarchy Process. New York: McGraw-Hill. 1980

[34] D. Gedvilaite: The assessment of sustainable development of a country's regions. Doctoral dissertation. Vilnius: Technika. 2019

[35] V. Podvezko: Agreement of expert estimates. Technological and Economic Development of Economy, Vol. 11, 2005, No. 2, pp. 101-107

[36] I. Ubarte: Multiple criteria decision support and recommender system for the assessment of healthy and safe homes in the built environment. Doctoral dissertation. Vilnius: Technika. 2017

[37] L. G. Jevlanov: Teoretical and practical aspects of Decision Making (in Russian) Moscow: Economics. 1984

[38] A. G. Tutygin, V. B. Korobof: Преимущества и недостатки метода анализа иерархий [Advantages and disadvantages of the analytic hierarchy process]. Естественные и точные науки [Natural and Exact Sciences] 122, 2010, pp. 108-115

[39] W. K. M. Brauers, R. Ginevicius, V. Podvezko: Regional development in Lithuania considering multiple objectives by the MOORA method. Technological and economic development of economy, Vol. 16, No. 4, 2010, pp. 613-640

A Deep-learned Type-3 Fuzzy System and Its Application in Modeling Problems

Man-Wen Tian[1], Ardashir Mohammadzadeh[2], Jafar Tavoosi[3], Saleh Mobayen[4], Jihad H. Asad[5], Oscar Castillo[6], Annámaria R. Várkonyi-Kóczy[7]

[1]National key project laboratory, Jiangxi University of Engineering, Xinyu 338000, China; E-mail: mwtian@ygu.edu.cn

[2]Department of Electrical Engineering, Faculty of Engineering University of Bonab, Bonab, Iran; Email: a.mzadeh@ubonab.ac.ir

[3]Department of Electrical Engineering, Faculty of Engineering, Ilam University, Ilam, Iran; Email: jtavoosi@aut.ac.ir

[4]Future Technology Research Center, National Yunlin University of Science and Technology, Douliu 64002, Taiwan; Email: mobayens@yuntech.edu.tw

[5]Department of Physics, Faculty of Applied Sciences, Palestine Technical University, Tulkarm, Palestine; Email: j.asad@ptuk.edu.ps

[6]Division of Graduate Studies and Research, Tijuana Institute of Technology, Tijuana, Mexico. Email: ocastillo@tectijuana.mx

[7]Institute of Software Design and Software Development, Óbuda University, Budapest, Hungary; Email: varkonyi-koczy@uni-obuda.hu
Department of Informatics, J. Selye University, Komarno, Slovakia

Abstract: The modeling problem is one of the important topics in engineering applications. In various applications, it is required to find a mathematical model to represent the relationship between output and the associated input variables. In this study, an approach on basis of a new deep learned type-3 (T3) fuzzy logic system (FLS) is introduced. The modeling of CO_2 solubility on basis of temperature, molality of NaCl, and pressure is considered as an application. The monitoring of carbon dioxide (CO_2) solubility in brine is one of the effective approaches in carbon capture and sequestration technique to reduce it in the atmosphere. A new hybrid learning method is presented to optimize the suggested model. The new adaptation laws are carry-out to tune the rule parameters and centers of membership functions (MFs). The values of horizontal slices and α- cuts are learned by the unscented Kalman filter (UKF). By the real-world experimental data sets, several statistical examinations, and comparison with conventional well-known fuzzy neural networks (NNs) and learning methods, the reliability and good performance of the suggested method are demonstrated. Also, the sensitivity of the input variables is analyzed by the use of the Sobol approach.

Keywords: Carbon dioxide solubility; Fuzzy logic systems; Learning algorithm; Estimation performance; Kalman filter

1 Introduction

Modeling problem is one of the important topics in engineering applications. In various applications, it is required to find a mathematical model to represent the relationship between output and associated input variables. In recent years, one of the main approaches that have been frequently reported in the literature to decrease carbon dioxide (CO_2) is carbon capture and sequestration (CCS). The solubility of CO_2 in brine is the basic factor in the CCS. Then forecasting the solubility of CO_2 with the desired accuracy is an important research topic [1].

For modeling and forecasting the solubility of CO_2, many approaches have been presented. For example, in [2], some experimental data in a special pressure and temperature is obtained, and then using Peng–Robinson equations a mathematical model is extracted. In [3], a thermodynamic model by the use of Soave–Redlich–Kwong equations is presented and its accuracy is discussed by some experimental data. The reliability of various existing models is discussed in [4], and the effect of pressure is studied. In [5], the development of scientific models for CO_2 is reviewed and the practical conditions to evaluate the capability of various models are discussed. In [6], by the use of VPT, CPA-SRK72, and PC-SAFT equations a model is presented to predict the solubility of CO_2 in brine and water, and the sensitivity of CO_2 solubility with respect to the temperature and pressure is investigated. In [7], some experimental data at high pressure for CO_2 solubility in brine is measured and then by Whitson equations, a model is developed. In [8], using Setschenow coefficients a developed model is presented and its performance is compared with other models and also the behavior of CO_2 solubility in versus of temperature is investigated. In [9], some new experimental data is presented and the Patel-Teja equation is taken to account to construct a model and its agreement with gathered data is investigated. In [10], by the staticanalytic method some new experimental data is extracted and two models are developed by the use of Robinson Cubic and Soreide and Whitson equations. In [11], the Soave–Redlich–Kwong equation is developed to obtain a model, and its superiority against the Peng–Robinson model is studied.

The machine learning techniques and intelligent systems such as FLS and neural networks have been successfully employed on various engineering problems such as complex problem applications [12], risk solving [13], prediction problems [14], optimization [15], susceptibility analysis [16], classification problems [17], control systems [18], among many others. However, rarely studies can be found in literature about modeling and forecasting CO_2 solubility by these techniques. For example, in [19], the least-square support-vector machine (LSSVM) is

proposed to estimate the CO_2 solubility, and its effectiveness is investigated in versus of optimized FLS by particle swarm optimization (PSO) technique. Similar to [19], the superiority of LSSVM against multi-layer perceptron (MLP) and radial basis function NN (RBF) is shown in [20]. It is declared that the R-squared value for LSSVM is 0.991 is versus of 0.964 and 0.916 for MLP and RBF, respectively. In [21], the Adaptive Boosting (AdaBoost) algorithm and NNs concept are combined to establish a model and its carnality is compared with LSSVMs. Similarly, in [22], the superiority of the AdaBoost technique is investigated by comparison with MLPs and LSSVMs. In [23], an Extra Trees model is developed and it is declared that the performance of the suggested method is better than FLS and NN approaches. In [24], two systems on basis of RBF and FLS are optimized by the genetic algorithm (GA) to estimate the CO_2 solubility. In [25], an FLS is optimized to approximate the CO_2 solubility and its accuracy and convergence velocity are studied. In [26], the FLS, NN, and self-organizing map techniques are combined to construct an intelligent model to predict CO_2 as a function of economic enhancement and energy consumption. In [27], an FLS model is developed, and by the use of Monte-Carlo method the sensitivity of the model with respect to pressure and temperature is investigated. In [28], an FLS is tuned by PSO algorithm and it is applied for CO_2 and methane solubility. In [29], the CO_2 viscosity is modeled by MLP and the parameters of MLP are optimized by the Levenberg-Marquardt (LM) algorithm and it is shown that the use of LM results in better accuracy in contrast to conjugate gradient algorithm. Similarly, in [30], an FLS is learned by various evolutionary optimization methods such as artificial bee colony (ABC), PSO, and GA, and the superiority of optimized FLS by PSO is shown. In [31], the effectiveness of the LSSVM in estimating the CO_2 solubility is studied and the influence of salinity, pressure, and temperature is analyzed. In [32], a model is developed by the use of GA and support vector machine (SVM), and its proficiency is examined in versus of NN-based approaches. In [33-36], decision making techniques are used for modeling problems in engineering applications. Recently it has been shown that the high-order FLSs are more effective than conventional NN-based approaches in modeling complicated nonlinear systems. However, it has not been used in CCS problem.

Motivated by the above literature review, in this paper, a new approach on basis of deep learned type-3 FLS is proposed to construct a model for CO_2 solubility estimation. For the first time, in addition to the rule parameters, the centers of MFs and level of horizontal slices are also learned. The effectiveness of the suggested method is examined by several statistical analyses and comparisons with conventional well-known methods. The main highlights are:

- For the first time, a deep learned T3-FLS is proposed.
- A hybrid learning method is presented such that, in addition to the rule parameters, the centers of MFs and level of horizontal slices are also learned.
- A new approach is presented to investigate the sensitivity of input variables.

- Several statistical analyses are provided to verify the effectiveness of the suggested T3-FLS.
- Some comparisons with various fuzzy neural networks and learning algorithms are provided to show the superiority of the suggested T3-FLS and hybrid learning method.

2 Proposed Type-3 FLS

2.1 General View

The type-3 FLS [37], is the generalization of the type-2 FLS that has more capacity to cope with uncertainties. A general view on the suggested T3-FLS is depicted in Figure 1. In T3-FLSs, as shown in Figure 2, the secondary membership function (MF) is also a type-2 MF. Then the upper and lower bounds of memberships are not constant in contrast to the type-2 MFs. This features cause that more level of uncertainties can be handled by type-3 MFs.

2.2 Structure

In this section, the process of computing is explained:

1) The inputs are *T*, *P*, and *M*, which represents temperature ($^{\circ}K$), pressure (bar), and molality of NaCl ($mol \cdot kg^{-1}$), respectively.

2) For each inputs *T*, *P*, and *M*, two membership functions (MFs) are considered. The MFs of inputs *T*, *P* and *M* are denoted as $\tilde{S}_T^1 - \tilde{S}_T^2$ and $S_P^1 - \tilde{S}_P^2$, respectively. Each MF is horizontally sliced into *n* levels as shown in Figure 2. As shown in Figure 3, for each input, the upper and lower memberships for horizontal slice level α_h, are computed. For input *T* the upper and lower memberships at horizontal splice level α_h are obtained as:

$$\overline{\mu}_{\tilde{S}_T^j|\overline{\alpha}_h}(T) = \exp\left(-\frac{\left(T - c_{\tilde{S}_T^j|\overline{\alpha}_h}\right)^2}{\overline{\sigma}^2_{\tilde{S}_T^j|\overline{\alpha}_h}}\right), \overline{\mu}_{\tilde{S}_T^j|\underline{\alpha}_h}(T) = \exp\left(-\frac{\left(T - c_{\tilde{S}_T^j|\underline{\alpha}_h}\right)^2}{\overline{\sigma}^2_{\tilde{S}_T^j|\underline{\alpha}_h}}\right) \quad (1)$$

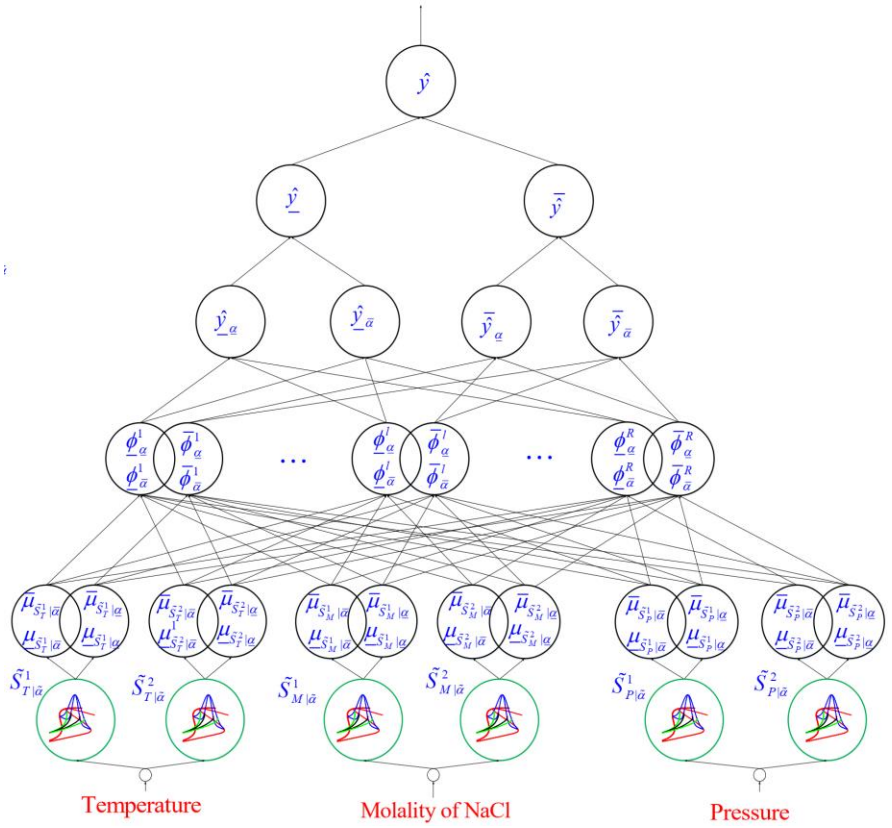


Figure 1
A general view on the suggested type-3 fuzzy logic system

$$\underline{\mu}_{\tilde{S}_T^j|\bar{\alpha}_h}(T)=\exp\left(-\frac{\left(T-c_{\tilde{S}_T^j|\bar{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_T^j|\bar{\alpha}_h}}\right),\underline{\mu}_{\tilde{S}_T^j|\underline{\alpha}_h}(T)=\exp\left(-\frac{\left(T-c_{\tilde{S}_T^j|\underline{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_T^j|\underline{\alpha}_h}}\right) \quad (2)$$

where, h = 1,...,n, j = 1,2, $c_{\tilde{S}_T^j|\bar{\alpha}_h}$ is the center of MF $\tilde{S}_T^j\,|\,\bar{\alpha}_h$, $\bar{\sigma}_{\tilde{S}_T^j|\bar{\alpha}_h}$, and $\underline{\sigma}_{\tilde{S}_T^j|\bar{\alpha}_h}$ are the upper and lower standard divisions for $\tilde{S}_T^j\,|\,\bar{\alpha}_h$. For input P, one has:

$$\bar{\mu}_{\tilde{S}_P^j|\bar{\alpha}_h}(P)=\exp\left(-\frac{\left(P-c_{\tilde{S}_P^j|\bar{\alpha}_h}\right)^2}{\bar{\sigma}^2_{\tilde{S}_P^j|\bar{\alpha}_h}}\right),\bar{\mu}_{\tilde{S}_P^j|\underline{\alpha}_h}(P)=\exp\left(-\frac{\left(P-c_{\tilde{S}_P^j|\underline{\alpha}_h}\right)^2}{\bar{\sigma}^2_{\tilde{S}_P^j|\underline{\alpha}_h}}\right) \quad (3)$$

$$\underline{\mu}_{\tilde{S}_P^j|\bar{\alpha}_h}(P)=\exp\left(-\frac{\left(P-c_{\tilde{S}_P^j|\bar{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_P^j|\bar{\alpha}_h}}\right),\underline{\mu}_{\tilde{S}_P^j|\underline{\alpha}_h}(P)=\exp\left(-\frac{\left(P-c_{\tilde{S}_P^j|\underline{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_P^j|\underline{\alpha}_h}}\right) \quad (4)$$

where, $h = 1,...,n$, $j = 1,2$, $c_{\tilde{S}_P^j|\bar{\alpha}_h}$ is the center of MF $\tilde{S}_P^j \mid \bar{\alpha}_h$, $\underline{\sigma}_{\tilde{S}_P^j|\bar{\alpha}_h}$ and $\underline{\sigma}_{\tilde{S}_P^j|\underline{\alpha}_h}$ are the upper and lower standard division for $\tilde{S}_P^j \mid \bar{\alpha}_h$. Similarly, for input M, one has:

$$\bar{\mu}_{\tilde{S}_M^j|\bar{\alpha}_h}=\exp\left(\frac{-\left(M-c_{\tilde{S}_M^j|\bar{\alpha}_h}\right)^2}{\bar{\sigma}^2_{\tilde{S}_M^j|\bar{\alpha}_h}}\right),\bar{\mu}_{\tilde{S}_M^j|\underline{\alpha}_h}=\exp\left(\frac{-\left(M-c_{\tilde{S}_M^j|\underline{\alpha}_h}\right)^2}{\bar{\sigma}^2_{\tilde{S}_M^j|\underline{\alpha}_h}}\right) \quad (5)$$

$$\underline{\mu}_{\tilde{S}_M^j|\bar{\alpha}_h}=\exp\left(-\frac{\left(M-c_{\tilde{S}_M^j|\bar{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_M^j|\bar{\alpha}_h}}\right),\underline{\mu}_{\tilde{S}_M^j|\underline{\alpha}_h}=\exp\left(-\frac{\left(M-c_{\tilde{S}_M^j|\underline{\alpha}_h}\right)^2}{\underline{\sigma}^2_{\tilde{S}_M^j|\underline{\alpha}_h}}\right) \quad (6)$$

where, $h = 1,...,n$, $j = 1,2$, $c_{\tilde{S}_P^j|\bar{\alpha}_h}$ is the center of MF $\tilde{S}_M^j \mid \bar{\alpha}_h$, $\bar{\sigma}_{\tilde{S}_M^j|\bar{\alpha}_h}$ and $\underline{\sigma}_{\tilde{S}_M^j|\bar{\alpha}_h}$ are the upper and lower standard division for $\tilde{S}_M^j \mid \bar{\alpha}_h$.

3) The upper rule firing at $\bar{\alpha}_h$ are computed as:

$$\bar{\varphi}^1_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\bar{\alpha}_h},\bar{\varphi}^2_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\bar{\alpha}_h} \quad (7)$$

$$\bar{\varphi}^3_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\bar{\alpha}_h},\bar{\varphi}^4_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\bar{\alpha}_h} \quad (8)$$

$$\bar{\varphi}^5_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\bar{\alpha}_h},\bar{\varphi}^6_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\bar{\alpha}_h} \quad (9)$$

$$\bar{\varphi}^7_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\bar{\alpha}_h},\bar{\varphi}^8_{\bar{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\bar{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\bar{\alpha}_h} \quad (10)$$

For upper rule firing at $\underline{\alpha}_h$, one has:

$$\bar{\varphi}^1_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\underline{\alpha}_h},\bar{\varphi}^2_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\underline{\alpha}_h} \quad (11)$$

$$\bar{\varphi}^3_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\underline{\alpha}_h},\bar{\varphi}^4_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\underline{\alpha}_h} \quad (12)$$

$$\bar{\varphi}^5_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\underline{\alpha}_h},\bar{\varphi}^6_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^1|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\underline{\alpha}_h} \quad (13)$$

$$\bar{\varphi}^7_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^1|\underline{\alpha}_h},\bar{\varphi}^8_{\underline{\alpha}_h}=\bar{\mu}_{\tilde{S}_T^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_P^2|\underline{\alpha}_h}\bar{\mu}_{\tilde{S}_M^2|\underline{\alpha}_h} \quad (14)$$

Similarly, the lower firing rules at upper and lower slices are computed as:

$$\underline{\varphi}^{1}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\bar{\alpha}_h}, \underline{\varphi}^{2}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\bar{\alpha}_h} \tag{15}$$

$$\underline{\varphi}^{3}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\bar{\alpha}_h}, \underline{\varphi}^{4}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\bar{\alpha}_h} \tag{16}$$

$$\underline{\varphi}^{5}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\bar{\alpha}_h}, \underline{\varphi}^{6}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\bar{\alpha}_h} \tag{17}$$

$$\underline{\varphi}^{7}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\bar{\alpha}_h}, \underline{\varphi}^{8}_{\bar{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\bar{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\bar{\alpha}_h} \tag{18}$$

$$\underline{\varphi}^{1}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\underline{\alpha}_h}, \underline{\varphi}^{2}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\underline{\alpha}_h} \tag{19}$$

$$\underline{\varphi}^{3}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\underline{\alpha}_h}, \underline{\varphi}^{4}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^1_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\underline{\alpha}_h} \tag{20}$$

$$\underline{\varphi}^{5}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\underline{\alpha}_h}, \underline{\varphi}^{6}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\underline{\alpha}_h} \tag{21}$$

$$\underline{\varphi}^{7}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^1_M|\underline{\alpha}_h}, \underline{\varphi}^{8}_{\underline{\alpha}_h} = \underline{\mu}_{\tilde{S}^2_T|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_P|\underline{\alpha}_h}\underline{\mu}_{\tilde{S}^2_M|\underline{\alpha}_h} \tag{22}$$

4) For the first type-reduction, the upper and lower of estimated output are computed as:

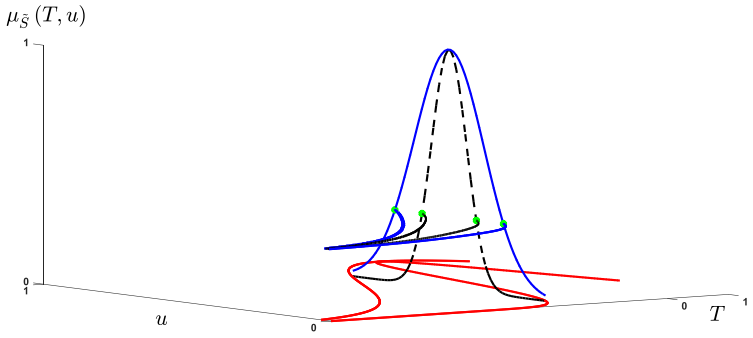


Figure 2
The horizontal slices of type-3 MF

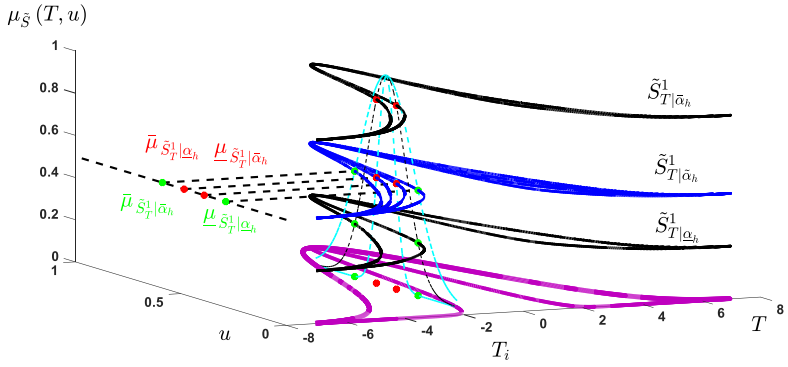


Figure 3

Representation of upper and lower memberships by two horizontal slices

$$\bar{\hat{y}}_{\bar{\alpha}_h} = \frac{\sum_{l=1}^{R} \bar{\varphi}^l_{\bar{\alpha}_h} \bar{\theta}_l}{\sum_{l=1}^{R} \left(\bar{\varphi}^l_{\bar{\alpha}_h} + \underline{\varphi}^l_{\bar{\alpha}_h} \right)}, \bar{\hat{y}}_{\underline{\alpha}_h} = \frac{\sum_{l=1}^{R} \bar{\varphi}^l_{\underline{\alpha}_h} \bar{\theta}_l}{\sum_{l=1}^{R} \left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h} \right)} \tag{23}$$

$$\underline{\hat{y}}_{\bar{\alpha}_h} = \frac{\sum_{l=1}^{R} \underline{\varphi}^l_{\bar{\alpha}_h} \underline{\theta}_l}{\sum_{l=1}^{R} \left(\bar{\varphi}^l_{\bar{\alpha}_h} + \underline{\varphi}^l_{\bar{\alpha}_h} \right)}, \underline{\hat{y}}_{\underline{\alpha}_h} = \frac{\sum_{l=1}^{R} \bar{\varphi}^l_{\underline{\alpha}_h} \underline{\theta}_l}{\sum_{l=1}^{R} \left(\underline{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h} \right)} \tag{24}$$

where, $R = 8$ is the number of rules, $\underline{\theta}_l$ and $\bar{\theta}_l$ are the lower and upper of $l-th$ rule parameters.

5) For the second type-reduction, one has:

$$\bar{\hat{y}} = \frac{\sum_{h=1}^{n} \bar{\alpha}_h \bar{\hat{y}}_{\bar{\alpha}_h}}{\sum_{h=1}^{n} \left(\bar{\alpha}_h + \underline{\alpha}_h \right)} + \frac{\sum_{h=1}^{n} \underline{\alpha}_h \bar{\hat{y}}_{\underline{\alpha}_h}}{\sum_{h=1}^{n} \left(\bar{\alpha}_h + \underline{\alpha}_h \right)}, \underline{\hat{y}} = \frac{\sum_{h=1}^{n} \bar{\alpha}_h \underline{\hat{y}}_{\bar{\alpha}_h}}{\sum_{h=1}^{n} \left(\bar{\alpha}_h + \underline{\alpha}_h \right)} + \frac{\sum_{h=1}^{n} \underline{\alpha}_h \underline{\hat{y}}_{\underline{\alpha}_h}}{\sum_{h=1}^{n} \left(\bar{\alpha}_h + \underline{\alpha}_h \right)} \tag{25}$$

6) The output $\hat{y}$ is the estimated solubility (mol · kg^{-1}) that is computed as:

$$\hat{y} = \frac{\bar{\hat{y}} + \underline{\hat{y}}}{2} \tag{26}$$

3 Learning Algorithm

In this section, the rule parameters, the centers of MFs, and the values of horizontal slices are tuned.

3.1 Tuning of Rule Parameters

The rule parameters are tuned by the EKF algorithm such that the following cost function are to be minimized:

$$J = \frac{1}{2}\left(y_d - \hat{y}\right)^2 \tag{27}$$

where, y_d is the desired solubility (mol·kg^{-1}) and $\hat{y}$ is the output of the suggested T3-FLS that represents the estimated solubility (mol·kg^{-1}). The tuning laws for the upper and lower rule parameters $\bar{\theta}$ and $\underline{\theta}$ are given as:

$$\bar{\theta}(t) = \bar{\theta}(t-1) + \bar{\pi}(t)\bar{\psi}(t)\left(y_d - \hat{y}\right), \underline{\theta}(t) = \underline{\theta}(t-1) + \underline{\pi}(t)\underline{\psi}(t)\left(y_d - \hat{y}\right) \tag{28}$$

where, $\bar{\pi}$ and $\underline{\pi}(t)$ are the corresponding covariance matrices for $\bar{\theta}$ an $\underline{\theta}$, respectively. The terms $\bar{\psi}(t)$ and $\underline{\psi}(t)$ are defined as:

$$\bar{\psi} = \left[\bar{\psi}_1, ..., \bar{\psi}_l, ..., \bar{\psi}_R\right]^T, \underline{\psi} = \left[\underline{\psi}_1, ..., \underline{\psi}_l, ..., \underline{\psi}_R\right]^T \tag{29}$$

where, $\bar{\psi}(t)$ and $\underline{\psi}(t)$ are:

$$\begin{aligned}\bar{\psi}_l &= \frac{\partial \hat{y}}{\partial \bar{\theta}_l} = \frac{\partial \hat{y}}{\partial \bar{\hat{y}}}\frac{\partial \bar{\hat{y}}}{\partial \bar{\theta}_l} = \frac{\partial \hat{y}}{\partial \bar{\hat{y}}}\frac{\partial \bar{\hat{y}}}{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}\frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\theta}_l} + \frac{\partial \hat{y}}{\partial \bar{\hat{y}}}\frac{\partial \bar{\hat{y}}}{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}\frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\theta}_l} \\ &= 0.5\frac{1}{\sum_{h=1}^{n}\left(\bar{\alpha}_h + \underline{\alpha}_h\right)}\sum_{h=1}^{n}\bar{\alpha}_h\frac{\bar{\varphi}^l_{\bar{\alpha}_h}}{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)} + 0.5\frac{1}{\sum_{h=1}^{n}\left(\bar{\alpha}_h + \underline{\alpha}_h\right)}\sum_{h=1}^{n}\underline{\alpha}_h\frac{\bar{\varphi}^l_{\underline{\alpha}_h}}{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)}\end{aligned} \tag{30}$$

$$\begin{aligned}\underline{\psi}_l &= \frac{\partial \hat{y}}{\partial \underline{\theta}_l} = \frac{\partial \hat{y}}{\partial \underline{\hat{y}}}\frac{\partial \underline{\hat{y}}}{\partial \underline{\theta}_l} = \frac{\partial \hat{y}}{\partial \underline{\hat{y}}}\frac{\partial \underline{\hat{y}}}{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}\frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\theta}_l} + \frac{\partial \hat{y}}{\partial \underline{\hat{y}}}\frac{\partial \underline{\hat{y}}}{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}\frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\theta}_l} \\ &= 0.5\frac{1}{\sum_{h=1}^{n}\left(\bar{\alpha}_h + \underline{\alpha}_h\right)}\sum_{h=1}^{n}\bar{\alpha}_h\frac{\underline{\varphi}^l_{\bar{\alpha}_h}}{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)} + 0.5\frac{1}{\sum_{h=1}^{n}\left(\bar{\alpha}_h + \underline{\alpha}_h\right)}\sum_{h=1}^{n}\underline{\alpha}_h\frac{\underline{\varphi}^l_{\underline{\alpha}_h}}{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)}\end{aligned} \tag{31}$$

3.2 Tuning of MF Parameters

For the antecedent parameters, the centers of MFs are tuned on basis of gradient descent method. Then the tuning laws are written as:

$$c_{\tilde{S}_T^j}(t) = c_{\tilde{S}_T^j}(t-1) - \gamma \frac{\partial J}{\partial c_{\tilde{S}_T^j}}, j = 1,2 \tag{32}$$

$$c_{\tilde{S}_P^j}(t) = c_{\tilde{S}_P^j}(t-1) - \gamma \frac{\partial J}{\partial c_{\tilde{S}_P^j}}, j = 1,2 \tag{33}$$

$$c_{\tilde{S}_M^j}(t) = c_{\tilde{S}_M^j}(t-1) - \gamma \frac{\partial J}{\partial c_{\tilde{S}_M^j}}, j = 1,2 \tag{34}$$

where, γ is the training rate. $\partial J / \partial c_{\tilde{S}_T^1}$ is obtained as follows:

$$\begin{aligned}
\frac{\partial J}{\partial c_{\tilde{S}_T^1}} &= \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \bar{\hat{y}}} \frac{\partial \bar{\hat{y}}}{\partial \bar{\hat{y}}_{\bar{\alpha}_h}} \left(\sum_{l=1}^{R} \bar{\zeta}_T^l \frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}_{\bar{\alpha}_h}^l} \frac{\partial \bar{\varphi}_{\bar{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} + \sum_{l=1}^{R} \underline{\zeta}_T^l \frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}_{\bar{\alpha}_h}^l} \frac{\partial \underline{\varphi}_{\bar{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} \right) + \\
&= \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \bar{\hat{y}}} \frac{\partial \bar{\hat{y}}}{\partial \bar{\hat{y}}_{\underline{\alpha}_h}} \left(\sum_{l=1}^{R} \bar{\zeta}_T^l \frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}_{\underline{\alpha}_h}^l} \frac{\partial \bar{\varphi}_{\underline{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} + \sum_{l=1}^{R} \underline{\zeta}_T^l \frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}_{\underline{\alpha}_h}^l} \frac{\partial \underline{\varphi}_{\underline{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} \right) + \\
&= \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \underline{\hat{y}}} \frac{\partial \underline{\hat{y}}}{\partial \underline{\hat{y}}_{\underline{\alpha}_h}} \left(\sum_{l=1}^{R} \bar{\zeta}_T^l \frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}_{\underline{\alpha}_h}^l} \frac{\partial \bar{\varphi}_{\underline{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} + \sum_{l=1}^{R} \underline{\zeta}_T^l \frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}_{\underline{\alpha}_h}^l} \frac{\partial \underline{\varphi}_{\underline{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} \right) + \\
&= \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \underline{\hat{y}}} \frac{\partial \underline{\hat{y}}}{\partial \underline{\hat{y}}_{\bar{\alpha}_h}} \left(\sum_{l=1}^{R} \bar{\zeta}_T^l \frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}_{\bar{\alpha}_h}^l} \frac{\partial \bar{\varphi}_{\bar{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} + \sum_{l=1}^{R} \underline{\zeta}_T^l \frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}_{\bar{\alpha}_h}^l} \frac{\partial \underline{\varphi}_{\bar{\alpha}_h}^l}{\partial c_{\tilde{S}_T^1}} \right)
\end{aligned} \tag{35}$$

where, $\bar{\zeta}_T^l$ represents the *l–th* element of vector $\bar{\zeta}_T$. The vector $\bar{\zeta}_T$ is defined as:

$$\bar{\zeta}_T = [1,1,1,1,0,0,0,0] \tag{36}$$

where, the elements of $\bar{\zeta}_T$ in the rules that include $c_{\tilde{S}_T^1}$, are one. The terms $\frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}_{\bar{\alpha}_h}^l}$, $\frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}_{\bar{\alpha}_h}^l}$, $\frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}_{\underline{\alpha}_h}^l}$, $\frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}_{\underline{\alpha}_h}^l}$, $\frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}_{\bar{\alpha}_h}^l}$, $\frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}_{\bar{\alpha}_h}^l}$, $\frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}_{\underline{\alpha}_h}^l}$ and $\frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}_{\underline{\alpha}_h}^l}$ are obtained as:

$$\frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}_{\bar{\alpha}_h}^l} = \bar{\theta}_l \frac{\sum_{l=1}^{R} \left(\bar{\varphi}_{\bar{\alpha}_h}^l + \underline{\varphi}_{\bar{\alpha}_h}^l \right) - \bar{\varphi}_{\bar{\alpha}_h}^l}{\left(\sum_{l=1}^{R} \left(\bar{\varphi}_{\bar{\alpha}_h}^l + \underline{\varphi}_{\bar{\alpha}_h}^l \right) \right)^2}, \frac{\partial \bar{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}_{\bar{\alpha}_h}^l} = \bar{\theta}_l \frac{-1}{\left(\sum_{l=1}^{R} \left(\bar{\varphi}_{\bar{\alpha}_h}^l + \underline{\varphi}_{\bar{\alpha}_h}^l \right) \right)^2} \tag{37}$$

$$\frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}^l_{\underline{\alpha}_h}} = \bar{\theta}_l \frac{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right) - \bar{\varphi}^l_{\underline{\alpha}_h}}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)\right)^2}, \frac{\partial \bar{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}^l_{\underline{\alpha}_h}} = \bar{\theta}_l \frac{-1}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)\right)^2} \tag{38}$$

$$\frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \bar{\varphi}^l_{\bar{\alpha}_h}} = \underline{\theta}_l \frac{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\bar{\alpha}_h} + \underline{\varphi}^l_{\bar{\alpha}_h}\right) - \bar{\varphi}^l_{\bar{\alpha}_h}}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\bar{\alpha}_h} + \underline{\varphi}^l_{\bar{\alpha}_h}\right)\right)^2}, \frac{\partial \underline{\hat{y}}_{\bar{\alpha}_h}}{\partial \underline{\varphi}^l_{\bar{\alpha}_h}} = \underline{\theta}_l \frac{-1}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\bar{\alpha}_h} + \underline{\varphi}^l_{\bar{\alpha}_h}\right)\right)^2} \tag{39}$$

$$\frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \bar{\varphi}^l_{\underline{\alpha}_h}} = \underline{\theta}_l \frac{\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right) - \bar{\varphi}^l_{\underline{\alpha}_h}}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)\right)^2}, \frac{\partial \underline{\hat{y}}_{\underline{\alpha}_h}}{\partial \underline{\varphi}^l_{\underline{\alpha}_h}} = \underline{\theta}_l \frac{-1}{\left(\sum_{l=1}^{R}\left(\bar{\varphi}^l_{\underline{\alpha}_h} + \underline{\varphi}^l_{\underline{\alpha}_h}\right)\right)^2} \tag{40}$$

For $\frac{\partial \bar{\varphi}^l_{\bar{\alpha}_h}}{\partial c_{\tilde{S}^1_T}}$, $\frac{\partial \bar{\varphi}^l_{\underline{\alpha}_h}}{\partial c_{\tilde{S}^1_T}}$, $\frac{\partial \underline{\varphi}^l_{\bar{\alpha}_h}}{\partial c_{\tilde{S}^1_T}}$ and $\frac{\partial \underline{\varphi}^l_{\underline{\alpha}_h}}{\partial c_{\tilde{S}^1_T}}$ one has:

$$\frac{\partial \bar{\varphi}^l_{\bar{\alpha}_h}}{\partial c_{\tilde{S}^1_T}} = \frac{2\left(T - c_{\tilde{S}^1_T|\bar{\alpha}_h}\right)}{\bar{\sigma}^2_{\tilde{S}^1_T|\bar{\alpha}_h}} \bar{\varphi}^l_{\bar{\alpha}_h}, \frac{\partial \bar{\varphi}^l_{\underline{\alpha}_h}}{\partial c_{\tilde{S}^1_T}} = \frac{2\left(T - c_{\tilde{S}^1_T|\underline{\alpha}_h}\right)}{\bar{\sigma}^2_{\tilde{S}^1_T|\underline{\alpha}_h}} \bar{\varphi}^l_{\underline{\alpha}_h} \tag{41}$$

$$\frac{\partial \underline{\varphi}^l_{\bar{\alpha}_h}}{\partial c_{\tilde{S}^1_T}} = \frac{2\left(T - c_{\tilde{S}^1_T|\bar{\alpha}_h}\right)}{\bar{\sigma}^2_{\tilde{S}^1_T|\bar{\alpha}_h}} \underline{\varphi}^l_{\bar{\alpha}_h}, \frac{\partial \underline{\varphi}^l_{\underline{\alpha}_h}}{\partial c_{\tilde{S}^1_T}} = \frac{2\left(T - c_{\tilde{S}^1_T|\underline{\alpha}_h}\right)}{\bar{\sigma}^2_{\tilde{S}^1_T|\underline{\alpha}_h}} \underline{\varphi}^l_{\underline{\alpha}_h} \tag{42}$$

The computation for terms $\partial J / \partial c_{\tilde{S}^2_T}$, $\partial J / \partial c_{\tilde{S}^1_M}$, $\partial J / \partial c_{\tilde{S}^2_M}$, $\partial J / \partial c_{\tilde{S}^1_P}$ and $\partial J / \partial c_{\tilde{S}^2_P}$, are the same as $\partial J / \partial c_{\tilde{S}^1_T}$, with difference that $\bar{\zeta}_T$ is repealed with $\underline{\zeta}_T$, $\bar{\zeta}_M$, $\underline{\zeta}_M$, $\bar{\zeta}_P$ and $\underline{\zeta}_P$, respectively. Also terms $c_{\tilde{S}^1_T|\bar{\alpha}_h}$, $c_{\tilde{S}^1_T|\underline{\alpha}_h}$, $\bar{\sigma}_{\tilde{S}^1_T|\bar{\alpha}_h}$ and $\bar{\sigma}^2_{\tilde{S}^1_T|\underline{\alpha}_h}$ should be replaced with the corresponding terms. The vectors $\underline{\zeta}_T$, $\bar{\zeta}_M$, $\underline{\zeta}_M$, $\bar{\zeta}_P$ and $\underline{\zeta}_P$ are defined as:

$$\underline{\zeta}_T = [0,0,0,0,1,1,1,1], \bar{\zeta}_M = [1,0,1,0,1,0,1,0] \tag{43}$$

$$\underline{\zeta}_M = [0,1,0,1,0,1,0,1], \bar{\zeta}_P = [1,1,0,0,1,1,0,0], \underline{\zeta}_P = [0,0,1,1,0,0,1,1] \tag{44}$$

3.3 Optimizing of Horizontal Slices

For the optimizing of horizontal slices level, the UKF algorithm is used. To apply UKF algorithm the state space of the suggested T3-FLS is written as follows:

$$\begin{aligned} \alpha(t+1) &= \alpha(t) + \nu(t) \\ \hat{y}(t+1) &= \text{T3-FLS}\left(u(t) \mid \theta(t) \mid \alpha(t)\right) + \upsilon(t) \end{aligned} \tag{45}$$

where, $v(t)$ and $\upsilon(t)$ represent noise with covariance v_p and υ_m, respectively. $u(t)$, $\theta(t)$ and $\alpha(t)$ are the vectors of input variables, consequent parameters and values of horizontal slices, respectively. The sigma points $\tilde{\alpha}$ are computed as:

$$\tilde{\alpha}_z = \tilde{\alpha}_z + \tilde{w}_z, z = 1,...,2n \tag{46}$$

where, n is the number of α-cuts and

$$\tilde{w}_z = \left(\sqrt{n\psi(t)}\right)^T, z = 1,...,n \tag{47}$$

$$\tilde{w}_{z+n} = -\left(\sqrt{n\psi(t)}\right)^T, z = 1,...,n \tag{48}$$

where, $\psi(t)$ is the covariance matrix. For each $\tilde{\alpha}(t)$ in (46), the output of T3-FLS are computed as:

$$\tilde{y}_z(t+1) = \text{T3-FLS}\left(u(t) \mid \theta(t) \mid \tilde{\alpha}_z(t)\right) \tag{49}$$

From (49), the average $\bar{y}$ is:

$$\bar{y} = \sum_{z=1}^{2n} \tilde{y}_z / (2n) \tag{50}$$

The cross-covariance $\psi_{\alpha y}$ is obtained as:

$$\psi_{\alpha y} = \frac{1}{2n} \sum_{z=1}^{2n} (\tilde{\alpha}_z - \hat{\alpha})\bar{y} \tag{51}$$

where, $\hat{\alpha}$ is computed as:

$$\hat{\alpha} = \frac{1}{2n} \sum_{z=1}^{2n} \tilde{\alpha}_z \tag{52}$$

Kalman gain is computed as:

$$K(t) = \psi_{\alpha y} \psi^{-1} \tag{53}$$

Finally, the vector of α-cuts are updated as:

$$\alpha(t+1) = \alpha(t) - K(t)\bar{y}(t) \tag{54}$$

5 Evaluation Indexes and Data Description

To examine the performance of the suggested method, 550 real-world data are collected from [7, 38]. The maximum of pressure, molality, temperature, and solubility are 1400, 6.14, 723.15, and 12.35, respectively. The minimum of pressure, molality, temperature, and solubility are 0.98, 0.016, 273.15, and 0.01, respectively. The data is normalized into the range [0,1]. 80% of data is randomly selected for training process and remains for testing.

To examine the capability of the suggested method, the following indexes are employed:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2},\mathrm{VAR}=\sum_{i=1}^{N}\left(y-\hat{y}_i\right)^2/\left(N-1\right) \tag{55}$$

$$\mathrm{TIC}=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}y_i^{\,2}}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i^{\,2}}},\mathrm{R}^2=1-\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}y_i-\hat{y}_i\right)^2} \tag{56}$$

where, N represents the number of data, y_i is the real solubility, $\hat{y}_i$ is the estimated solubility and RMSE, TIC and VAF are root mean square error, Theil's inequality coefficient, and variance account for, respectively.

6 Simulation

In this section, the performance of the suggested T3-FLS and hybrid learning algorithm is examined.

6.1 Results for Testing Data

To examine the estimation performance of the suggested T3-FLS and learning algorithm several statistical analyses were presented. Estimation performance for test and training data are shown in Figure 4. It is seen that the estimated solubility is well converged to the real one and the estimation error is at a desired and logical level.

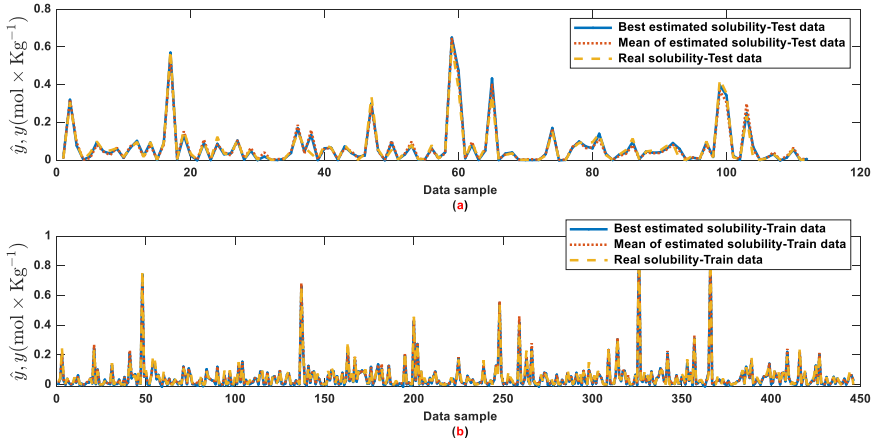


Figure 4

(a): Estimation performance for testing data ; (b): Estimation performance for training data

The absolute error for testing data and the values of RMSE, VAR, and TIC for testing data are given in Figure 5. One can see that the maximum and of the absolute error in the worst state is less than 2.

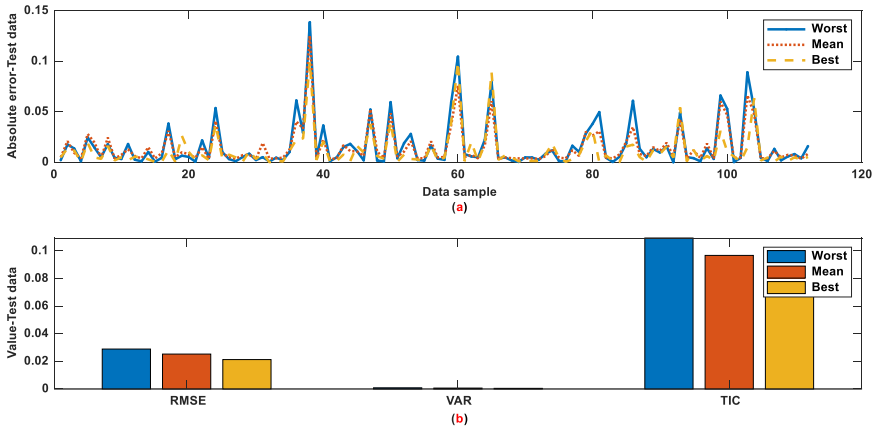


Figure 5

(a): Absolute error for testing data; (b): The values of RMSE, VAR and TIC for testing data

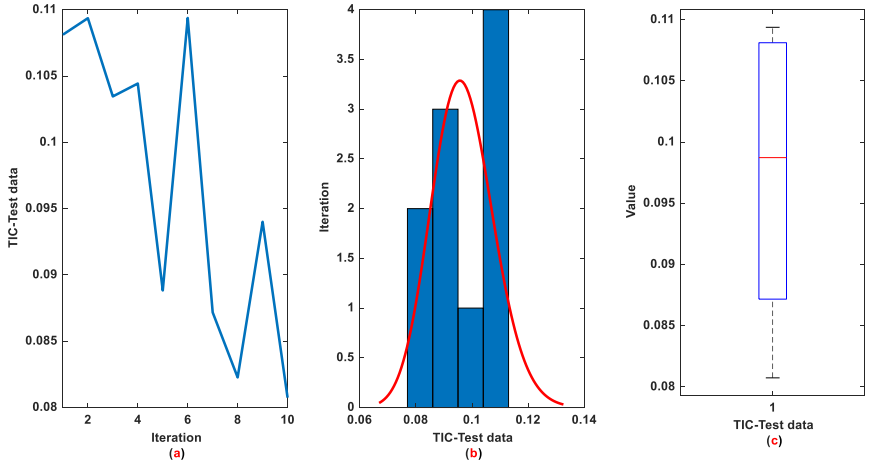


Figure 6

(a): The value of TIC for testing data; (b): Histogram diagram for TIC for testing data; (c): Box plot of TIC for testing data

The value of TIC, the histogram plot for TIC, and Box plot of TIC for testing data are depicted in Figure 6. It is seen that the mean of TIC is about 0.08, and the maximum of TIC is less than 0.09. The value of RMSE at each iteration, the histogram plot of RMSE, and Box plot of RMSE for testing data are shown in Figure 5. The mean of RMSE is about 0.25 and the maximum of RMSE is less than 0.26. The value of VAR at each iteration, the histogram plot of VAR, and the Box plot of VAR for testing data are given in Figure 8. It is seen that the value of VAR is small enough and the results in various iterations are close to each other. The values of R2 in iterations are shown in Figure 9. As it is seen the average of R2 is greater than 0.95, which represents a good correlation.

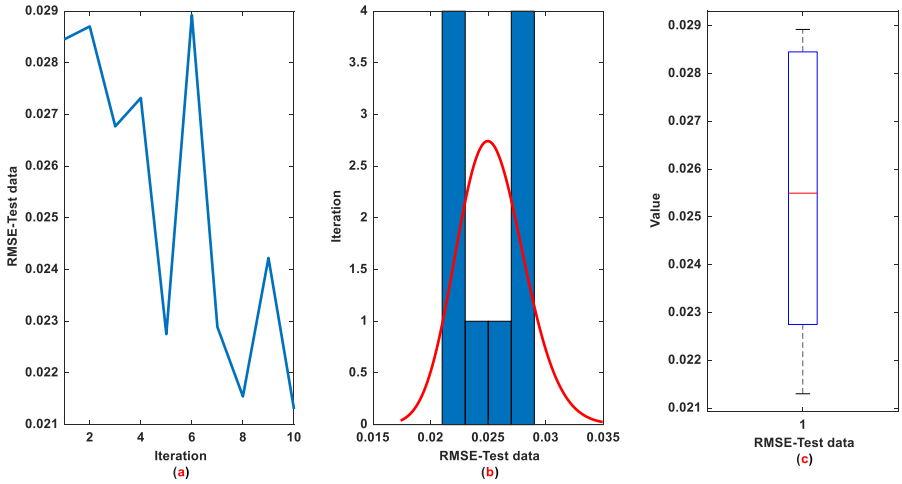


Figure 7

(a): The value of RMSE at each iteration for testing data; (b): Histogram plot of RMSE for testing data; (c): Box plot of RMSE for testing data

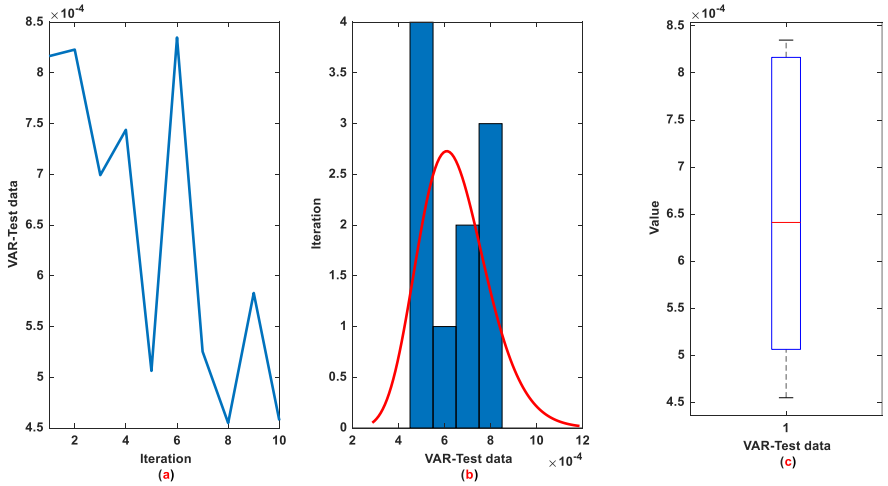


Figure 8

(a): The value of VAR at each iteration for testing data; (b): Histogram plot of VAR for testing data; (c): Box plot of VAR for testing data

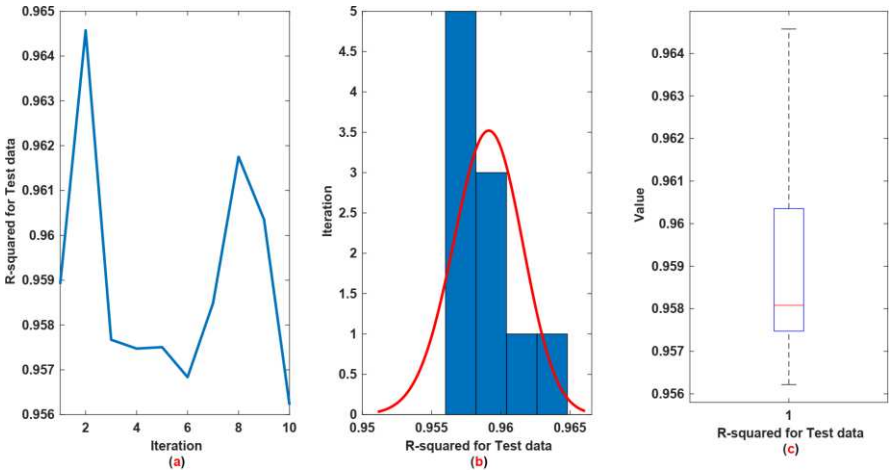


Figure 9
(a): The value of R2 at each iteration for testing data; (b): Histogram plot of R2 for testing data; (c): Box plot of R2 for testing data

6.3 Sensitivity Analysis

For sensitivity analysis, the Sobol method is employed. In this approach, the variance of the solubility is decomposed on its input variance. The sensitivity index is a value between 0 and 1. The larger index represents the higher influence. The values of the first-order sensitivity index for the suggested T3-FLS model are given in Table 1. As it is seen the most effective variable of solubility is the pressure.

Table 1
Sensitivity analysis by Sobol approach

Structure	Temperature	Pressure	Molality of N
First-Order Sensitivity Effect	0.0396	0.2291	0.1256

6.4 Comparison and Discussion

A comparison of RMSE with various neuro-fuzzy systems is given in Table 2. The performance of the suggested method is compared with the MLP, RBF, type-1 FLS (T1-FLS), and type-2 FLS (T2-FLS). It is observed that the suggested T3-FLS results in better solubility perdition performance. To better show the accommodation between model output and measured data, the cross plot for train and testing data is depicted in Figure 10. It is seen that most of the training and testing data are near the unit slope line. This plot verifies the good estimation

performance and well accommodation. Furthermore, a comparison with other learning methods is given in Table 3. The proposed hybrid learning system is compared with GA, PSO, and ABC algorithms that are frequently used in literature for optimization of models of CO2 solubility. It is seen that the suggested hybrid learning method results in better accuracy.

Table 2

Comparison of RMSE with various neuro-fuzzy systems

Structure	MLP	RBF	T1-FLS	T2-FLS	T3-FLS
RMSE	0.34	0.33	0.31	0.27	0.23
R^2	0.81	0.82	0.83	0.85	0.95

Table 3

Comparison of RMSE with various learning algorithms

Learning Method	PSO	GA	ABC	Proposed
RMSE	0.31	0.30	0.28	0.23

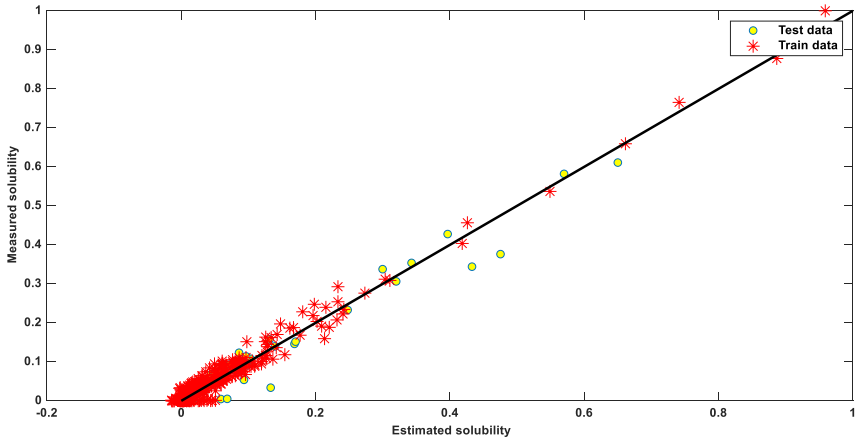


Figure 10

Cross plot for train and testing data

Conclusions

In this study, a new approach on basis of type-3 FLSs is proposed to construct a model between solubility and affective variables such as temperature, pressure, and molality of NaCl. The parameters of the suggested model are tuned by a hybrid learning method using EKF and UKF. The rule and centers of membership functions are optimized by EKF and the level of horizontal slices are tuned by the UKF algorithm. Several statistical analyses such as plotting the trajectories of TIC, VAR, RMSE, Box plot, R-squared, and histogram plot are provided to

demonstrate the effectiveness and reliability of the suggested method. It is shown that the estimation performance in various initial conditions does not change significantly. Also, two comparisons with conventional well-known structures and well-known learning algorithms demonstrate the superiority of the presented T3-FLS and learning algorithm. Furthermore, sensitivity analysis by the Sobol approach is provided to show the most effective input variable of T3-FLS.

References

[1] M. M. Balas, Seven Passive Greenhouse Synergies, Acta Polytechnica Hungarica 11 (4) (2014) 199-210

[2] E. Ali, M. K. Hadj-Kali, S. Mulyono, I. Alnashef, A. Fakeeha, F. Mjalli, A. Hayyan, Solubility of co2 in deep eutectic solvents: experiments and modelling using the peng–robinson equation of state, Chemical Engineering Research and Design 92 (10) (2014) 1898-1906, https://doi.org/10.1016/j.cherd.2014.02.004

[3] N. I. Diamantonis, G. C. Boulougouris, D. M. Tsangaris, M. J. El Kadi, H. Saadawi, S. Negahban, I. G. Economou, Thermodynamic and transport property models for carbon capture and sequestration (ccs) processes with emphasis on co2 transport, Chemical engineering research and design 91 (10) (2013) 1793-1806, https://doi.org/10.1016/j.cherd.2013.06.017

[4] D. Allen, B. Strazisar, Y. Soong, S. Hedges, Modeling carbon dioxide sequestration in saline aquifers: Significance of elevated pressures and salinities, Fuel processing technology 86 (14-15) (2005) 1569-1580, https://doi.org/10.1016/j.fuproc.2005.01.004

[5] A. Riaz, Y. Cinar, Carbon dioxide sequestration in saline formations: Part I review of the modeling of solubility trapping, Journal of Petroleum Science and Engineering 124 (2014) 367-380, https://doi. org/10.1016/j.petrol.2014.07.024

[6] A. Hassanpouryouzband, M. V. Farahani, J. Yang, B. Tohidi, E. Chuvilin, V. Istomin, B. Bukhanov, Solubility of flue gas or carbon dioxide-nitrogen gas mixtures in water and aqueous solutions of salts: Experimental measurement and thermodynamic modeling, Industrial & Engineering Chemistry Research 58 (8) (2019) 3377-3394, https://doi.org/10.1021/acs.iecr.8b04352

[7] W. Yan, S. Huang, E. H. Stenby, Measurement and modeling of co2 solubility in nacl brine and co2– saturated nacl brine density, International Journal of Greenhouse Gas Control 5 (6) (2011) 1460-1477, https://doi.org/10.1016/j.ijggc.2011.08.004

[8] H. Zhao, R. Dilmore, D. E. Allen, S. W. Hedges, Y. Soong, S. N. Lvov, Measurement and modeling of co2 solubility in natural and synthetic formation brines for co2 sequestration, Environmental science & technology 49 (3) (2015) 1972-1980, https://doi.org/10.1021/es505550a

[9] A. Chapoy, A. Mohammadi, A. Chareton, B. Tohidi, D. Richon, Measurement and modeling of gas solubility and literature review of the properties for the carbon dioxide- water system, Industrial & engineering chemistry research 43 (7) (2004) 1794-1802, https://doi.org/10.1021/ie034232t

[10] S. Chabab, P. Theveneau,´ J. Corvisier, C. Coquelet, P. Paricaud, C. ouriez, E. El Ahmar, Thermodynamic study of the co2–h2o–nacl system: easurements of co2 solubility and modeling of phase equilibria using soreide and whitson, electrolyte cpa and sit models, International Journal of reenhouse Gas Control 91 (2019) 102825, https://doi.org/10.1016/j.ijggc.2019.102825

[11] G. Sodeifian, N. S. Ardestani, S. A. Sajadian, Solubility measurement of a pigment (phthalocyanine green) in supercritical carbon dioxide: Experimental correlations and thermodynamic modeling, Fluid Phase Equilibria 494 (2019) 61-73, https://doi.org/10.1016/j.fluid.2019.04.024

[12] A. Pejic,´ P. S. Molcer, Predictive machine learning approach for complex problem solving process data mining, Acta Polytechnica Hungarica 18 (1)

[13] A. Di Noia, A. Martino, P. Montanari, A. Rizzi, Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction, Soft Computing 24 (6) (2020) 4393-4406

[14] G. Grmanova,´ P. Laurinec, V. Rozinajova,´ A. B. Ezzeddine, M. Lucka,´ P. Lacko, P. Vrablecova,´ P. Navrat,´ Incremental ensemble learning for electricity load forecasting, Acta Polytechnica Hungarica 13 (2) (2016) 97-117

[15] T. T. Ngoc, C. M. T. Le Van Dai, C. M. Thuyen, Support vector regression based on grid search method of hyperparameters for load forecasting, Acta Polytechnica Hungarica 18 (2) (2021) 143-158

[16] H. Moayedi, A. Osouli, D. T. Bui, L. Kok Foong, H. Nguyen, B. Kalantar, Two novel neural-evolutionary predictive techniques of dragonfly algorithm (da) and biogeography-based optimization (bbo) for landslide susceptibility analysis, Geomatics, Natural Hazards and Risk 10 (1) (2019) 2429-2453

[17] T. Stepisnik,˘ D. Kocev, S. Dzeroski,˘ Option predictive clustering trees for multi-label classification, Acta Polytechnica Hungarica 17 (10)

[18] A. Khamis, J. Meng, J. Wang, A. T. Azar, E. Prestes, H. Li, I. A. Hameed, A. Takacs,´ I. J. Rudas, T. Haidegger, Robotics and intelligent systems against a pandemic, Acta Polytechnica Hungarica 18 (5)

[19] M. Raji, A. Dashti, P. Amani, A. H. Mohammadi, Efficient estimation of co2 solubility in aqueous salt solutions, Journal of Molecular Liquids 283 (2019) 804-815, https://doi.org/10.1016/j.molliq.2019.02.090

[20] A. Baghban, A. Bahadori, A. H. Mohammadi, A. Behbahaninia, Prediction of co2 loading capacities of aqueous solutions of absorbents using different computational schemes, International Journal of Greenhouse Gas Control 57 (2017) 143-161, https://doi.org/10.1016/j.ijggc.2016.12.010

[21] H. Saghafi, M. Arabloo, Modeling of co2 solubility in mea, dea, tea, and mdea aqueous solutions using adaboost-decision tree and artificial neural network, International Journal of Greenhouse Gas Control 58 (2017) 256-265, https://doi.org/10.1016/j.ijggc.2016.12.014

[22] H. Saghafi, M. M. Ghiasi, A. H. Mohammadi, Co2 capture with aqueous solution of sodium glycinate: Modeling using an ensemble method, International Journal of Greenhouse Gas Control 62 (2017) 23-30, https://doi.org/10.1016/j.ijggc.2017.03.029

[23] H. Yarveicy, H. Saghafi, M. M. Ghiasi, A. H. Mohammadi, Decision tree-based modeling of co2 equilibrium absorption in different aqueous solutions of absorbents, Environmental Progress & Sustainable Energy 38 (s1) (2019) S441-S448, https://doi.org/10.1002/ep.13128

[24] S.-A. Hoseinpour, A. Barati-Harooni, P. Nadali, A. Mohebbi, A. Najafi-Marghmaleki, A. Tatar, A. Bahadori, Accurate model based on artificial intelligence for prediction of carbon dioxide solubility in aqueous tetra-n-butylammonium bromide solutions, Journal of Chemometrics 32 (2) (2018) e2956, https://doi.org/10.1002/cem.2956

[25] A. Baghban, H. Rajabi, N. Jamshidi, ANFIS modeling of carbon dioxide capture from gas stream emissions in the petrochemical production units, Petroleum Science and Technology 35 (6) (2017) 625-631, https://doi.org/10.1080/10916466.2016.1273241

[26] A. Mardani, H. Liao, M. Nilashi, M. Alrasheedi, F. Cavallaro, A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques, Journal of Cleaner Production 275 (2020) 122942, https://doi.org/10.1016/j.jclepro.2020.122942

[27] A. Daryasafar, A. Keykhosravi, K. Shahbazi, Modeling co2 wettability behavior at the interface of brine/co2/mineral: Application to co2 geo-sequestration, Journal of Cleaner Production 239 (2019) 118101, https://doi.org/10.1016/j.jclepro.2019.118101

[28] Y. Wang, Toward modeling of solubility of carbon dioxide, methane, and nitrogen in liquid dibenzyl toluene (ldt) using rigorous technique, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects (2019) 1-10, https://doi.org/10.1080/15567036.2019.1694100

[29] M. N. Amar, M. A. Ghriga, H. Ouaer, M. E. A. B. Seghier, B. T. Pham, P. Ø. Andersen, Modeling viscosity of co2 at high temperature and pressure

conditions, Journal of Natural Gas Science and Engineering 77 (2020) 103271, https://doi.org/10.1016/j.jngse.2020.103271

[30] A. Bemani, A. Baghban, A. Mosavi, Estimating co2-brine diffusivity using hybrid models of anfis and evolutionary algorithms, Engineering Applications of Computational Fluid Mechanics 14 (1) (2020) 818-834, https://doi.org/10.1080/19942060.2020.1774422

[31] M. Afkhami Karaei, B. Honarvar, A. Azdarpour, E. Mohammadian, On prediction of carbon dioxide solubility in aqueous systems of nacl using lssvm algorithm, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects (2019) 1-10, https://doi.org/10.1080/15567036.2019.1651789

[32] Q. Feng, R. Cui, S. Wang, J. Zhang, Z. Jiang, Estimation of co2 diffusivity in brine by use of the genetic algorithm and mixed kernels-based support vector machine model, Journal of Energy Resources Technology 141 (4), https://doi.org/10.1115/1.4041724

[33] Wang H, Liu DH, Yu L, Chen GJ, Yan HP, Tian MW, Huang HL, Yan SR, Wang LN, Li WH. Analysis on the reading guidance to classics in the teaching of specialized courses in tourism undergraduate education in China. The International Journal of Electrical Engineering & Education. 2020 Jun 30:0020720920931073

[34] Tian M, Yan S, Tian X. Discrete approximate iterative method for fuzzy investment portfolio based on transaction cost threshold constraint. Open Physics. 2019 Jan 1;17(1):41-7

[35] Tian MW, Wang S, Yan SR. Damage Assessment of Marine Pollution Disasters to Biological Resources Based on Numerical Models. Acta Microscopica. 2019 Apr 1;28(3)

[36] Tian MW, Yan SR, Tian XX, Liu JA. Research on image recognition method of bank financing bill based on binary tree decision. Journal of Visual Communication and Image Representation. 2019 Apr 1;60:123-8

[37] A. Mohammadzadeh, M. H. Sabzalian, W. Zhang, An interval type-3 fuzzy system and a new online fractional-order learning algorithm: theory and practice, IEEE Transactions on Fuzzy Systems 28 (2019) 1940-1950, https://doi.org/10.1109/TFUZZ.2019.2928509

[38] M. S. Aktar, M. De, S. Maity, S. K. Mazumder, M. Maiti, Green 4d transportation problems with breakable incompatible items under type-2 fuzzy-random environment, Journal of Cleaner Production 275 (2020) 122376

The Arc Teeth Semi-rolled Cylindrical Gear Meshing Geometry

Vladimir Syzrantsev, Ksenia Syzrantseva

Industrial University of Tyumen, 38 Volodarskogo St., 625000 Tyumen, Russia
E-mail: syzrantsevvn@tyuiu.ru, syzrantsevakv@tyuiu.ru

Abstract: Arc teeth semi-rolled cylindrical gears in the conditions of unbraced machine bodies have a higher load capacity, durability and reliability as well as the ability to compensate for the shaft axis twist angle by self-adjustment of one of the wheels compared to straight and helical teeth gears. In the article, the study object is the arc teeth semi-rolled cylindrical gear. The wheel arc teeth are cut using the single division method without generating with the cutting head, the generating surface of which is a straight circular cone. The gear arc tooth flank is an envelope of the wheel teeth flanks family at a given relative movement. The semi-rolled version of arc teeth cylindrical gears allows to significantly simplify the technological process of cutting wheels and producing gears with large gear ratios. Mathematical models of wheel and pinion arc teeth flanks forming processes have been built for a semi-rolled cylindrical gear. Dependences for calculating principal curvatures of the wheel and pinion arc teeth flanks have been obtained. An algorithm and a program for calculating the coordinates of the points of the active path of action in arc teeth meshing and principal relative curvatures at these points have been developed. The problem of determining the size of the contact pattern in the studied gear has been solved. Calculations to estimate variations in the position of the active paths of action and the sizes of the contact pattern with varying the wheel and pinion axes twist angle and variations of the principal relative curvatures in the longitudinal and profile directions of the arc teeth at the points of active paths of action have been performed.

Keywords: semi-rolled cylindrical gear; arc teeth; teeth flank curvatures; lines of action; principal relative curvatures; contact pattern

1 Introduction

Cylindrical straight, helical and double helical teeth gears are the basis of most modern machines and mechanisms. The theoretical meshing contact in these gears occurs along the line. In the transmissions of tractors, locomotives, coal-mining and other energy-intensive machines, cylindrical gears are mounted on projecting shafts. In gear operations, deformations of shafts and bodies lead to misalignment of teeth in meshing, their edge contact, and a multiple reduction in the service life

of gears. An effective way to increase the loading capacity and durability of cylindrical gears under the specified operating conditions is to use arc teeth (Figure 1). Arc teeth meshing may provide a linear, locally linear and point-to-point contact of the tooth flanks. Optimization of the geometry of these gears makes it possible to skip shifting the contact area to the edge of the tooth in the gear operation under teeth misalignment. The increased bending strength of arc teeth and the possibility of compensating for the twist (misalignment) angle of the teeth by self-adjustment of one of the transmission elements indicate the effectiveness of using arc teeth cylindrical gears in the drives of modern energy-saturated machines.

Figure 1
Arc teeth cylindrical gears

Currently, a number of methods are proposed for cutting generating cylindrical gear wheel arc teeth, which differ in the used tools and forming movements [1, 2, 3, 4]. The helix angle of the arc tooth in its midsection is equal to zero. Arc teeth cylindrical gear meshing in the proposed methods is based on the counterpart rack. In [1, 5, 6], variants of cutting cylindrical wheel arc teeth with circular cutting heads by means of generating with single division on CNC machines are considered.

The methods worked out for cutting spiral teeth of generating bevel gear wheels are the basis of most of the above methods [7, 8]. For mass production of bevel gears, it has become possible to reduce the cost of their manufacture by switching to a semi-rolled gear option [9, 10]. The wheel arc teeth of such a gear are cut with a cutting head without generating at a single division. Gleason specialists have developed the FORMATE and HELIXFORM methods in this area. Despite the more difficult task of finding the optimal geometry [9, 10], the technology of manufacturing such bevel gears is more advanced.

The analysis of the works related to the study of geometry [1, 2, 3, 4], contact and bending strength and durability [1, 11, 12] of arc teeth gears shows that all of them are dedicated to generating cylindrical gears. The models of forming arc teeth flanks for semi-rolled cylindrical gears are considered in the works [13, 14]. Prior to this work, the issues of calculating the geometric characteristics of the arc teeth contact (the coordinates of the contact points in different meshing phases, the

principal relative curvatures at the contact points, and the size of the contact pattern) in meshing of the semi-rolled cylindrical gear have not been considered. The noted geometric characteristics of arc teeth meshing are necessary for developing a calculation method for contact loading and load-bearing capacity by the value of the transmitted torque of arc teeth cylindrical gears.

2 Forming the Semi-rolled Cylindrical Gear Wheel Arc Tooth Usable Flank

According to the methods of forming semi-rolled bevel gears [9, 10], in the manufacture of semi-rolled cylindrical gears, the arc teeth of the wheel are cut by the single division method without generating with a cutting head, the generating surface of which is a straight circular cone. In this case, the usable flank of the wheel arc tooth will also be the surface of the straight circular cone. Forming the usable flank of the pinion arc tooth is implemented on the basis of the generating wheel. The flank of the arc pinion tooth is the envelope of the flank of the wheel tooth at a given relative movement of the wheel and pinion in the transmission. Modern four-axes CNC machines allow implementing this method of cutting arc pinion teeth [5, 6]. The transmission formed according to this scheme, in the absence of errors in the relative position of the pinion and wheel, is matched (theoretically accurate).

We describe the generating surface of the cutting head (straight circular cone) in coordinate system $S_p(x_p, y_{p,} z_p)$ rigidly connected to it (Figure 2), as follows [13, 14]:

$$x_p = \cos\vartheta\left(u \cdot \sin\alpha_0 - r_{g2}\right);\ y_p = u \cdot \cos\alpha_0\ ;\ z_p = \sin\vartheta\left(u \cdot \sin\alpha_0 - r_{g2}\right), \qquad (1)$$

where: u, ϑ are linear and angular parameters of the generating surface; α_0 is a basic profile angle; r_{g2} is a calculated radius of the cutting head rotating around axis y_p of coordinate system $S_p(x_p, y_{p,} z_p)$.

Taking into account the way of forming the usable surface of the wheel tooth, radius-vector $\bar{r}_p(x_p, y_{p,} z_p)$ is to be written into the coordinate system $S_2(x_2, y_2\, z_2)$, rigidly connected to the wheel:

$$\tilde{r}_2 = \tilde{A}_{2,p} \cdot \tilde{r}_p, \qquad (2)$$

where: $\tilde{A}_{2,p}$ is a fourth-order matrix describing the transition from coordinate system $S_p(x_p, y_{p,} z_p)$ to system $S_2(x_2, y_2\, z_2)$, the elements of which are

determined in accordance with Figure 3; $\tilde{r}_2$, $\tilde{r}_p$ are columns matrixes made up of vector radii coordinates $\bar{r}_2$ and $\bar{r}_p$ respectively.

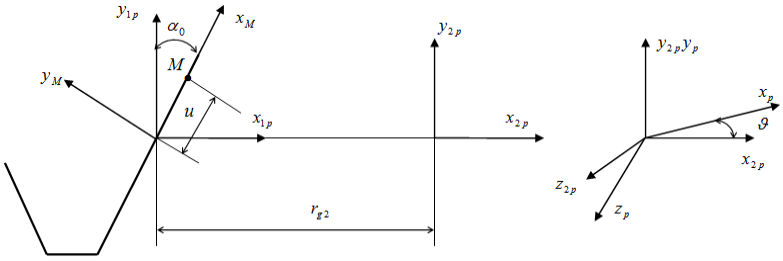


Figure 2
Cutting head generating surface

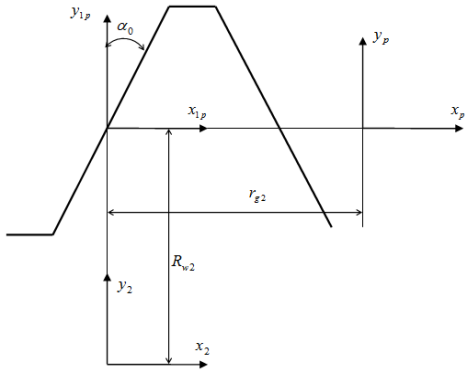


Figure 3
Coordinate systems to define elements of matrix $\tilde{A}_{2,p}$

Matrix $\tilde{A}_{2,p}$ has the following form:

$$\tilde{A}_{2,p} = \begin{Vmatrix} 1 & 0 & 0 & r_{g2} \\ 0 & 1 & 0 & R_{w2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix} \tag{3}$$

where: R_{w2} is the radius of the pitch circle.

We find the projections of the wheel radius-vector $\bar{r}_2(x_2, y_{2,} z_2)$ based on (1) and (3), opening equation (2):

$$x_2 = \cos\vartheta\left(u \cdot \sin\alpha_0 - r_{g2}\right) + r_{g2};\ y_2 = u \cdot \cos\alpha_0 + R_{w2};\ z_2 = \sin\vartheta\left(u \cdot \sin\alpha_0 - r_{g2}\right). \tag{4}$$

Projections m_{2x}, m_{2y}, m_{2z} of the flank normal unitary vector (4) have the form:

$$m_{2x} = \cos\alpha_0 \cdot \cos\vartheta; \quad m_{2y} = -\sin\alpha_0; \quad m_{2z} = \cos\alpha_0 \cdot \sin\vartheta. \tag{5}$$

Expressions (4) and (5) describe radius-vector $\bar{r}_2(x_2, y_2, z_2)$ and normal unitary vector $\bar{m}_2(m_{2x}, m_{2y}, m_{2z})$ of the usable flank of the convex side of the wheel arc tooth in coordinate system $S_2(x_2, y_2\, z_2)$.

From the theory of spatial gearing by meshing, it is known [7], that to determine the law of stress distribution over the contact area in meshing during the torque transmission, it is required to be able to calculate the gap at the point of contact of the teeth usable flanks. This gap is accurate to the values of the second order of smallness in main directions to determine the principal relative curvatures calculated at the point of contact of the tooth flanks. To calculate the principal curvatures of the arc teeth flanks and principal relative curvatures in meshing of the semi-rolled cylindrical gear, we use the methods developed in the theory of spatial gearing [7]. Following these methods, for the arc wheel teeth flanks, curvatures (k) are determined based on the ratio:

$$d\bar{m}_2 = -k \cdot d\bar{r}_2, \tag{6}$$

where: $\bar{m}_2$ is the normal unitary vector of the arc wheel tooth flank (5); $\bar{r}_2$ is a radius vector of the arc wheel tooth flank (4).

Differentials of vectors $\bar{m}_2$ and $\bar{r}_2$ by parameters u and ϑ have the form:

$$d\bar{m}_2 = \bar{m}_{2u}du + \bar{m}_{2\vartheta}d\vartheta; \; d\bar{r}_2 = \bar{r}_{2u}du + \bar{r}_{2\vartheta}d\vartheta. \tag{7}$$

Here, the indices u and ϑ indicate the partial derivative by parameters u and ϑ, respectively.

Based on dependencies (7), we present expression (6) in the form:

$$\bar{m}_{2u}du + \bar{m}_{2\vartheta}d\vartheta = -k\left(\bar{r}_{2u}du + \bar{r}_{2\vartheta}d\vartheta\right). \tag{8}$$

This vector equation is equivalent to two scalar equations since vectors $d\bar{r}_2$ and $d\bar{m}_2$ lie in the tangent plane. Projecting these vectors onto axes x_2 and z_2 of coordinate systems $S_2(x_2, y_2\, z_2)$ rigidly connected to the wheel, we get two scalar equations:

$$m_{2xu}du + m_{2x\vartheta}d\vartheta = -k\left(x_{2u}du + x_{2\vartheta}d\vartheta\right); m_{2zu}du + m_{2z\vartheta}d\vartheta = -k\left(z_{2u}du + z_{2\vartheta}d\vartheta\right). \tag{9}$$

The system of two linear equations (9) is transformed to a quadratic equation, referred to k, of the following form:

$$k^2 \cdot A_2 + k \cdot B_2 + C_2 = 0, \tag{10}$$

where: $A_2 = z_{2u}x_{2\vartheta} - z_{2\vartheta}x_{2u}$; $B_2 = z_{2u}m_{2x\vartheta} + x_{2\vartheta}m_{2zu} - m_{2xu}z_{2\vartheta} - m_{2z\vartheta}x_{2u}$; $C_2 = m_{2zu}m_{2x\vartheta} - m_{2z\vartheta}m_{2xu}$, indices u and ϑ of projections on the axis of coordinate system $S_2(x_2, y_2\, z_2)$ of vectors $\bar{r}_2$ and $\overline{m}_2$ mean partial derivatives by parameters u and ϑ.

Differentiating expressions (4) and (5) by parameters u and ϑ, we define:

$x_{2u} = \cos\vartheta\sin\alpha_0$; $z_{2u} = \sin\vartheta\sin\alpha_0$; $x_{2\vartheta} = -\sin\vartheta\left(u\cdot\sin\alpha_0 - r_{g2}\right)$; $z_{2\vartheta} = \cos\vartheta\left(u\cdot\sin\alpha_0 - r_{g2}\right)$; $m_{2xu} = 0$; $m_{2x\vartheta} = -\cos\alpha_0\sin\vartheta$; $m_{2zu} = 0$; $m_{2z\vartheta} = \cos\alpha_0\cos\vartheta$.

Substituting these expressions in equation (10), after solving it, we obtain the dependences for the principal curvatures of the arc wheel flank:

$$k_{21} = -\cos\alpha_0 / \left(u\cdot\sin\alpha_0 - r_{g2}\right) \text{ and } k_{22} = 0. \tag{11}$$

At the calculated point of the wheel tooth flank ($u = 0$), formulas (11) are simplified:

$$k_{21} = \cos\alpha_0 / r_{g2},\ \ k_{22} = 0. \tag{12}$$

3 Forming the Semi-rolled Cylindrical Gearing Gear Arc Tooth Usable Flank

To determine the usable flank of the concave side of the pinion arc tooth, we use the fact that it is a one-parameter envelope of the family of wheel tooth usable flanks in a given relative motion - the rotation of the pinion and wheel with constant gear ratio $i^* = z_2^* / z_1^* = const$; z_1^*, z_2^* are the numbers of the pinion and wheel teeth.

Using the methods of the spatial meshing theory [7, 13, 14, 15], we write the equation of the pinion tooth usable flank in the form:

$$\tilde{r}_1(u, \vartheta, \varphi_2) = \tilde{A}_{1,2}(\varphi_2)\tilde{r}_2(u, \vartheta);\ \ f(u, \vartheta, \varphi_2) = 0. \tag{13}$$

Here: $\tilde{A}_{1,2}(\varphi_2)$ is a fourth-order matrix describing the transition from coordinate system $S_2(x_2, y_2\, z_2)$ to coordinate system $S_1(x_1, y_1\, z_1)$ rigidly connected to the pinion (Figure 4); φ_2 is the angle of the wheel rotation when forming the pinion tooth flank, associated with the angle of its rotation φ_1 through gear ratio i^*:

$$\varphi_1 = i^*\cdot\varphi_2 = (z_2^*\cdot\varphi_2) / z_1^*. \tag{14}$$

The meshing equation [7, 13] is the last written in (13).

Using Figure 4, we define elements a_{ij}, $i=\overline{1,4}$; $j=\overline{1,4}$ of matrix $\tilde{A}_{1,2}(\varphi_2)$:

$$a_{11}=\cos(\varphi_1+\varphi_2);\ a_{12}=\sin(\varphi_1+\varphi_2);\ a_{13}=0;\ a_{14}=-a_{ws}\sin\varphi_1;$$
$$a_{21}=-\sin(\varphi_1+\varphi_2);\ a_{22}=\cos(\varphi_1+\varphi_2);\ a_{23}=0;\ a_{24}=-a_{ws}\cos\varphi_1; \quad (15)$$
$$a_{31}=a_{32}=a_{34}=a_{41}=a_{42}=a_{43}=0;\ a_{33}=a_{44}=1,$$

where: a_{ws} - is the center distance in the machine meshing of the pinion and the generating wheel.

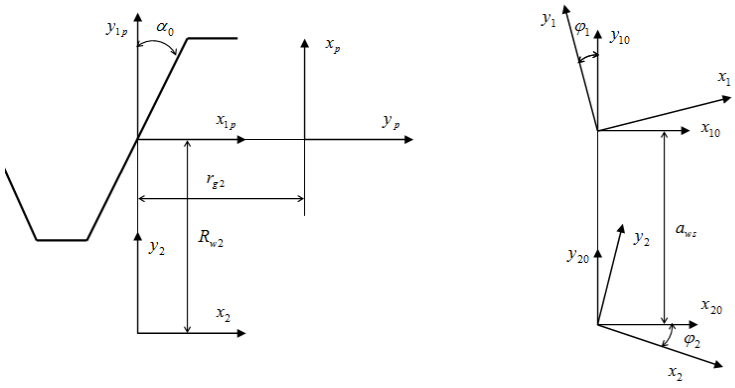


Figure 4
Coordinate systems to define elements of matrix $\tilde{A}_{1,2}(\varphi_2)$

Opening matrix equation (13) with respect to (4) and (15), we find the expressions for radius-vector $\bar{r}_1(x_1, y_{1,} z_1)$ projection of the usable flank of the pinion arc tooth:

$$x_1=A\cdot\cos(\varphi_1+\varphi_2)+B\cdot\sin(\varphi_1+\varphi_2)-a_{ws}\cdot\sin\varphi_1;$$
$$y_1=-A\cdot\sin(\varphi_1+\varphi_2)+B\cdot\cos(\varphi_1+\varphi_2)-a_{ws}\cdot\cos\varphi_1; \quad (16)$$
$$z_1=\sin\vartheta\left(u\cdot\sin\alpha_0-r_{g1}\right),$$

where: $A=\cos\vartheta\left(u\cdot\sin\alpha_0-r_{g1}\right)+r_{g1}$; $B=u\cdot\cos\alpha_0+R_{2w}$.

To obtain the meshing equation we use the method described in [13, 14], according to which, the meshing equation is written in the form:

$$f(u,\vartheta,\varphi_2)=\bar{V}_\varphi\cdot\bar{m}_2=V_{\varphi x}\cdot m_{2x}+V_{\varphi y}\cdot m_{2y}+V_{\varphi z}\cdot m_{2z}=0, \quad (17)$$

where $\bar{V}_\varphi$ is the vector analog of the relative speed, $\bar{m}_2$ is the normal unitary vector of the generating wheel arc tooth flank (5).

Projections $V_{\varphi x}$, $V_{\varphi y}$, $V_{\varphi z}$ of radius-vector $\overline{V}_{\varphi}$ are calculated according to the expression [13, 14]:

$$\tilde{V}_{\varphi} = \tilde{C}_{\varphi} \cdot \tilde{r}_2 . \tag{18}$$

Here: $\tilde{C}_{\varphi} = \tilde{A}_{1,2}^{-1} \cdot \dfrac{d\tilde{A}_{1,2}}{d\varphi_2}$ is the fourth-order matrix of the relative speed analog; $\tilde{V}_{\varphi}$ is a column matrix composed of projections $V_{\varphi x}$, $V_{\varphi y}$, $V_{\varphi z}$.

After differentiating elements (15) of matrix $\tilde{A}_{1,2}(\varphi_2)$ by φ_2 with respect to (14) based on the expression for matrix $\tilde{C}_{\varphi}$ (18), we obtain the following formulas for its elements c_{ij}, $j = \overline{1,4}$:

$$\begin{aligned} &c_{11} = 0; \quad c_{12} = (1 + i^*); \quad c_{13} = 0; \quad c_{14} = -a_{ws} \cdot i^* \cdot \cos\varphi_2; \\ &c_{21} = -(1 + i^*); \quad c_{22} = 0; \quad c_{23} = 0; \quad c_{24} = -a_{ws} \cdot i^* \cdot \sin\varphi_2; \\ &c_{31} = c_{32} = c_{33} = c_{44} = 0; \quad c_{41} = c_{42} = c_{43} = c_{44} = 0. \end{aligned} \tag{19}$$

Having the dependencies (19) and opening the matrix product (18) taking into account (4), we define expressions for radius-vector $\overline{V}_{\varphi}$ projections:

$$V_{\varphi x} = (1 + i^*) \cdot B - i^* \cdot a_{ws} \cdot \cos\varphi_2; \; V_{\varphi y} = (1 + i^*) \cdot B - i^* \cdot a_{ws} \cdot \sin\varphi_2; \; V_{\varphi z} = 0, \tag{20}$$

substituting which in equation (17) jointly with expressions (5) and (16), we obtain the meshing equation in the following form:

$$\begin{aligned} &f(u, \vartheta, \varphi_2) = u \cdot \cos\vartheta \cdot \left(1 + i^*\right) + \cos\alpha_0 \cdot \cos\vartheta \cdot \left[R_{2w} \cdot \left(1 + i^*\right) - i^* \cdot a_{ws} \cdot \cos\varphi_2\right] + \\ &+ \sin\alpha_0 \cdot \left[\left(1 + i^*\right) \cdot r_{g1} \cdot \left(1 - \cos\vartheta\right) - i^* \cdot a_{ws} \cdot \sin\varphi_2\right] = 0. \end{aligned} \tag{21}$$

This equation can be presented as analytically resolved or with respect to parameter u:

$$u = -\frac{\cos\alpha_0 \cos\vartheta \left[R_{2w}\left(1 + i^*\right) - i^* a_{ws} \cos\varphi_2\right] + \sin\alpha_0 \left[\left(1 + i^*\right) r_{g1}\left(1 - \cos\vartheta\right) - i^* a_{ws} \sin\varphi_2\right]}{(1 + i^*)\cos\vartheta} \tag{22}$$

or with respect to the parameter φ_2:

$$\varphi_2 = \arcsin\left(-C_u / \sqrt{A_u^2 + B_u^2}\right) - \xi, \tag{23}$$

where: $A_u = -i^* \cdot a_{ws} \cdot \sin\alpha_0$; $B_u = -i^* \cdot a_{ws} \cdot \cos\alpha_0 \cdot \cos\vartheta$; $\xi = \arcsin\left(B_u / \sqrt{A_u^2 + B_u^2}\right)$;

$$C_u = u \cdot \cos\vartheta \cdot \left(1 + i^*\right) + \cos\alpha_0 \cos\vartheta \cdot R_{w2}\left(1 + i^*\right) + \sin\alpha_0 \left(1 + i^*\right) \cdot r_{g1} \cdot \left(1 - \cos\vartheta\right).$$

Dependences (16) jointly with equation (21) fully describe the geometry of the usable flank of the concave side of the arc pinion tooth (13).

Projections, m_{1x}, m_{1y}, m_{1z}, of the normal unitary vector of the usable flank of the arc pinion tooth are determined based on the matrix equation:

$$\tilde{m}_1(u,\vartheta,\varphi_2) = \tilde{A}_{1,2}(\varphi_2)\tilde{m}_2(u,\vartheta)\,. \tag{24}$$

Opening the expression (24) based on dependences (15) and (5), we find:

$$\begin{aligned} m_{1x} &= \cos(\varphi_1+\varphi_2)\cos\alpha_0\cos\vartheta - \sin\alpha_0\sin(\varphi_1+\varphi_2)\,; \\ m_{1y} &= -\sin(\varphi_1+\varphi_2)\cos\alpha_0\cos\vartheta - \sin\alpha_0\cos(\varphi_1+\varphi_2)\,; \\ m_{1z} &= \cos\alpha_0\sin\vartheta\,. \end{aligned} \tag{25}$$

We obtain formulas for calculating principal curvatures of the pinion arc tooth flank the way we did previously for the wheel arc tooth flank. To solve this problem, we use dependence (6) in the form:

$$d\overline{m}_1 = -k\cdot d\overline{r}_1, \tag{26}$$

where: $\overline{m}_1$ is the normal unitary vector of the arc pinion tooth flank (25); $\overline{r}_1$ is the radius-vector of the arc pinion tooth flank (16). The flank of the pinion tooth is the envelope of the family of wheel tooth flanks and depends on three parameters: u, ϑ and φ_2, associated by the meshing equation (21). In this case, the differentials of vectors $\overline{m}_1$ and $\overline{r}_1$ by parameters u, ϑ and φ_2 have the form:

$$d\overline{m}_1 = \overline{m}_{1u}du + \overline{m}_{1\vartheta}d\vartheta + \overline{m}_{1\varphi}d\varphi\,;\; d\overline{r}_1 = \overline{r}_{1u}du + \overline{r}_{1\vartheta}d\vartheta + \overline{r}_{1\varphi}d\varphi\,, \tag{27}$$

and dependence (26) takes the form:

$$\overline{m}_{1u}du + \overline{m}_{1\vartheta}d\vartheta + \overline{m}_{1\varphi}d\varphi = -k\left(\overline{r}_{1u}du + \overline{r}_{1\vartheta}d\vartheta + \overline{r}_{1\varphi}\right). \tag{28}$$

Vectors $d\overline{r}_1$ and $d\overline{m}_1$ lie in the tangent plane. We project them on axes x_1 and z_1 of coordinate system $S_1(x_1, y_1 z_1)$ rigidly connected to the pinion, and take into account that parameters u, ϑ and φ_2 are associated by the meshing equation (21). As a result, we obtain the following system of three scalar equations:

$$\begin{aligned} &m_{1xu}du + m_{1x\vartheta}d\vartheta + m_{1x\varphi}d\varphi = -k\left(x_{1u}du + x_{1\vartheta}d\vartheta + x_{1\varphi}d\varphi\right); \\ &m_{z1u}du + m_{1z\vartheta}d\vartheta + m_{1z\varphi}d\varphi = -k\left(z_{1u}du + z_{1\vartheta}d\vartheta + z_{1\varphi}d\varphi\right); \\ &f_u du + f_\vartheta d\vartheta + f_\varphi d\varphi = 0\,, \end{aligned} \tag{29}$$

where: f_u, f_ϑ, f_φ are partial derivatives of the meshing equation (21) by parameters u, ϑ and φ_2.

Based on the third equation of the system (29), we have $d\varphi = -\left(f_u du + f_\vartheta d\vartheta\right)/ f_\varphi$, after which the system (29) is transformed to the form:

$$f_\varphi(m_{1xu} + k \cdot x_{1u})du + f_\varphi(m_{1x\vartheta} + k \cdot x_{1\vartheta}) - (m_{1x\varphi} + k \cdot x_{1\varphi})(f_u du + f_\vartheta d\vartheta) = 0;$$
$$f_\varphi(m_{1zu} + k \cdot z_{1u})du + f_\varphi(m_{1z\vartheta} + k \cdot z_{1\vartheta}) - (m_{1z\varphi} + k \cdot z_{1\varphi})(f_u du + f_\vartheta d\vartheta) = 0. \quad (30)$$

Following the theory of differential geometry, to determine the values of principal curvatures, it is necessary to equate the determinant of the system of equations (30) to zero

$$\begin{vmatrix} f_\varphi(m_{1xu} + k \cdot x_{1u}) - f_u\left(m_{1x\varphi} + k \cdot x_{1\varphi}\right); & f_\varphi(m_{1x\vartheta} + k \cdot x_{1\vartheta}) - f_\vartheta(m_{1x\varphi} + k \cdot x_{1\varphi}); \\ f_\varphi(m_{1zu} + k \cdot z_{1u}) - f_u(m_{1z\varphi} + k \cdot z_{1\varphi}); & f_\varphi(m_{1z\vartheta} + k \cdot z_{1\vartheta}) - f_\vartheta(m_{1z\varphi} + k \cdot z_{1\varphi}); \end{vmatrix} = 0 \quad (31)$$

and solve the resulting quadratic equation relative to k.

After the transformations, equation (31) has the following form:

$$w_1 \cdot k^2 + (w_2 + w_3) \cdot k + w_4 = 0, \quad (32)$$

where:

$$w_1 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ x_{1u} & x_{1\vartheta} & x_{1\varphi} \\ z_{1u} & z_{1\vartheta} & z_{1\varphi} \end{vmatrix}; \quad w_2 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ m_{1xu} & m_{1x\vartheta} & m_{1x\varphi} \\ z_{1u} & z_{1\vartheta} & z_{1\varphi} \end{vmatrix};$$

$$w_3 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ x_{1u} & x_{1\vartheta} & x_{1\varphi} \\ m_{1zu} & m_{1z\vartheta} & m_{1z\varphi} \end{vmatrix}; \quad w_4 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ m_{1xu} & m_{1x\vartheta} & m_{1x\varphi} \\ m_{1zu} & m_{1z\vartheta} & m_{1z\varphi} \end{vmatrix}. \quad (33)$$

Similar formulas can be obtained using the normal unitary vector and radius-vector of the generating surface, in this case, it is the radius-vector of the generating wheel surface (4) and its normal unitary vector (5). In this case, the calculations are simplified, and the coefficients of equation (32) have the form:

$$w_1 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ x_{2u} & x_{2\vartheta} & V_{\varphi x} \\ z_{2u} & z_{2\vartheta} & V_{\varphi z} \end{vmatrix}; \quad w_2 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ m_{2xu} & m_{2x\vartheta} & w_{\varphi x} \\ z_{2u} & z_{2\vartheta} & V_{\varphi z} \end{vmatrix};$$

$$w_3 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ x_{2u} & x_{2\vartheta} & V_{\varphi x} \\ m_{2zu} & m_{2z\vartheta} & w_{\varphi z} \end{vmatrix}; \quad w_4 = \begin{vmatrix} f_u & f_\vartheta & f_\varphi \\ m_{2xu} & m_{2x\vartheta} & w_{\varphi x} \\ m_{2zu} & m_{2z\vartheta} & w_{\varphi z} \end{vmatrix}. \quad (34)$$

Here: f_u, f_ϑ, f_φ are partial derivatives of the meshing equation (21) by parameters u, ϑ and φ_2; x_{2u}, $x_{2\vartheta}$, $z_{2u}, z_{2\vartheta}$ are partial derivatives by parameters u and ϑ of the radius-vector projections (4) on axes x_2 and z_2 of coordinate system $S_2(x_2, y_2, z_2)$; $V_{\varphi x}$ and $V_{\varphi z}$ are vector-analog projections (20) of the relative speed on axes x_2 and z_2; $w_{\varphi x}$, $w_{\varphi z}$ are projections of the vector-analog of the angular velocity on axes x_2 and z_2; $m_{2xu}, m_{2x\vartheta}$, m_{2zu}, $m_{2z\vartheta}$ are partial derivatives by parameters u and ϑ of projections of the normal unitary vector (5) of the surface (4).

Projections $w_{\varphi x}$ and $w_{\varphi z}$ are calculated based on the same expression (18):

$$\tilde{w}_\varphi = \tilde{C}_\varphi \cdot \tilde{m}_2 = \tilde{A}_{1,2}^{-1} \cdot \frac{d\tilde{A}_{1,2}}{d\varphi_2} \cdot \tilde{m}_2 \tag{35}$$

opening which on the basis of (19) and (5), we obtain:

$$w_{\varphi x} = -(1+i^*)\sin\alpha_0;\ w_{\varphi y} = -(1+i^*)\cos\alpha_0\cos\vartheta;\ w_{\varphi z} = 0. \tag{36}$$

The dependences for calculating the remaining elements of the determinants (34) have the form:

$$f_u = (1+i^*)\cos\vartheta;\ f_\varphi = i^* \cdot a_{ws}(\cos\alpha_0\cos\vartheta\sin\varphi_2 - \sin\alpha_0\cos\varphi_2);$$

$$f_\vartheta = -u(1+i^*)\sin\vartheta - \cos\alpha_0\sin\vartheta\left[R_{w2}\left(1+i^*\right) - i^* \cdot a_{ws}\cos\varphi_2\right] + (1+i^*)r_{g1}\sin\alpha_0\sin\vartheta;$$

$$x_{2u} = \cos\vartheta\sin\alpha_0;\quad x_{2\vartheta} = -\sin\vartheta\left(u\sin\alpha_0 - r_{g2}\right);\quad m_{2xu} = 0; \tag{37}$$

$$m_{2x\vartheta} = -\cos\alpha_0\sin\vartheta;\quad z_{2u} = \sin\vartheta\sin\alpha_0;\quad z_{2\vartheta} = \cos\vartheta(u\sin\alpha_0 - r_{g2});\quad m_{2zu} = 0;$$
$$m_{2z\vartheta} = \cos\alpha_0\cos\vartheta;\ V_{\varphi x} = (1+i^*)(u\cos\alpha_0 + R_{w2}) - i^* \cdot a_{ws} \cdot \cos\varphi_2;\quad V_{\varphi z} = 0.$$

We open the expressions (34) with respect to (37) and find:

$$w_1 = f_\vartheta \cdot V_{\varphi x} \cdot z_{2u} + f_\varphi\left(x_{2u} \cdot z_{2\vartheta} - x_{2\vartheta} \cdot z_{2u}\right) - f_u \cdot V_{\varphi x} \cdot z_{2\vartheta};$$
$$w_2 = f_\vartheta \cdot w_{\varphi x} \cdot z_{2u} - f_\varphi \cdot m_{2x\vartheta} \cdot z_{2u} - f_u \cdot w_{\varphi x} \cdot z_{2\vartheta}; \tag{38}$$
$$w_3 = m_{2z\vartheta}(f_\varphi \cdot x_{2u} - f_u \cdot V_{\varphi x});\quad w_4 = -f_u \cdot m_{2z\vartheta} \cdot w_{\varphi x}.$$

As a result, we solve the equation (32) based on (38) and obtain the following expression for calculation of principal curvatures of the pinion tooth flank:

the curvature by the length of the tooth is

$$k_{11} = \left[-\left(w_2 + w_3\right) - \sqrt{\left(w_2 + w_3\right)^2 - 4 \cdot w_1 \cdot w_4}\right] / \left(2 \cdot w_1\right). \tag{39}$$

the curvature by the profile of the tooth is

$$k_{12} = \left[-\left(w_2 + w_3\right) + \sqrt{\left(w_2 + w_3\right)^2 - 4 \cdot w_1 \cdot w_4} \right] / \left(2 \cdot w_1\right). \quad (40)$$

At the calculated point of the pinion tooth flank: $u = 0$, $\vartheta = 0$ and $\varphi_2 = \varphi_1 = 0$. For these values the formulas (39) and (40) are simplified and have the form:

$$k_{11} = -\cos\alpha_0 / r_{g2}; \quad k_{12} = -\left(1 + i^*\right)^2 / \left(i^* \cdot a_{ws} \cdot \sin\alpha_0\right). \quad (41)$$

4 Calculation of Contact Points Coordinates of Wheel and Pinion Arc Teeth Flanks in Meshing

Meshing semi-rolled cylindrical gear arc teeth is matched in the absence of errors of manufacturing and the wheel and pinion relative position, based on the pinion teeth forming method. Teeth meshing occurs in linear contact conditions. To exclude the edge contact that occurs due to errors in manufacturing teeth, it is localized in the longitudinal direction of the tooth. The gear remains matched, and the contact of the active flanks of the pinion and wheel teeth occurs in the tooth transverse midsection where the helix angle is zero. In the case when the gear operates under errors in the relative position of the wheel and pinion, their meshing becomes approximate with the teeth point contact.

The problem of determining the coordinates of the contact points of the active flanks of wheel and pinion arc teeth, installed with errors in their relative position, is an inverse problem of the meshing theory [7, 9, 14]. Knowing the coordinates of the contact points is required to solve the problem of calculating the contact loading of the gear. To solve the inverse problem, expressions of the radii-vectors and normal unitary vectors of pinion and wheel arc teeth flanks are required.

Based on the dependences (4), (5), the projections of radius-vector $\bar{r}_2^2(u_2, \vartheta_2)$ and normal unitary vector $\bar{m}_2^2(\vartheta_2)$ of the arc wheel tooth flank in coordinate system $S_2(x_2, y_2 z_2)$ (superscript "2" in the designation of vectors) rigidly connected to the wheel, have the form:

$$x_2(u_2, \vartheta_2) = \cos\vartheta_2\left(u_2 \cdot \sin\alpha_0 - r_{g2}\right) + r_{g2}; \quad y_2(u_2) = u_2 \cdot \cos\alpha_0 + R_{w2}; \quad (42)$$

$$z_2(u_2, \vartheta_2) = \sin\vartheta_2\left(u_2 \cdot \sin\alpha_0 - r_{g2}\right),$$

$$m_{2x}(\vartheta_2) = \cos\alpha_0 \cdot \cos\vartheta_2; \quad m_{2y} = -\sin\alpha_0; \quad m_{2z}(\vartheta_2) = \cos\alpha_0 \cdot \sin\vartheta_2. \quad (43)$$

where: u_2, ϑ_2 are linear and angular parameters of the wheel tooth flank.

Projections of radius vector $\bar{r}_1^1(u_1, \vartheta_1, \varphi_1)$ and normal unitary vector $\overline{m}_1^1(\vartheta_1, \varphi_1)$ of the pinion tooth flank in coordinate system $S_1(x_1, y_{1,} z_1)$ rigidly connected to the pinion, taking into account the formulas (16), (21) and (25), are described by the following expressions:

$$\begin{aligned} x_1(u_1, \vartheta_1, \varphi_1) &= A \cdot \cos(\varphi_1 + \varphi_2) + B \cdot \sin(\varphi_1 + \varphi_2) - a_{ws} \cdot \sin\varphi_1; \\ y_1(u_1, \vartheta_1, \varphi_1) &= -A \cdot \sin(\varphi_1 + \varphi_2) + B \cdot \cos(\varphi_1 + \varphi_2) - a_{ws} \cdot \cos\varphi_1; \end{aligned} \tag{44}$$

$$z_1(u_1, \vartheta_1) = \sin\vartheta_1\left(u_1 \cdot \sin\alpha_0 - r_{g1}\right),$$

$$\begin{aligned} f(u_1, \vartheta_1, \varphi_1) &= u_1 \cdot \cos\vartheta_1 \cdot (1+i) + \cos\alpha_0 \cdot \cos\vartheta_1 \cdot \left[R_{2w} \cdot (1+i) - i \cdot a_{ws} \cdot \cos\varphi_2\right] + \\ &+ \sin\alpha_0 \cdot \left[(1+i) \cdot r_{g1} \cdot (1-\cos\vartheta_1) - i \cdot a_{ws} \cdot \sin\varphi_2\right] = 0, \end{aligned} \tag{45}$$

$$z_2(u_2, \vartheta_2) = \sin\vartheta_2\left(u_2 \cdot \sin\alpha_0 - r_{g2}\right),$$

$$\begin{aligned} m_{1x}(\vartheta_1, \varphi_1) &= \cos(\varphi_1 + \varphi_2)\cos\alpha_0 \cos\vartheta_1 - \sin\alpha_0 \sin(\varphi_1 + \varphi_2); \\ m_{1y}(\vartheta_1, \varphi_1) &= -\sin(\varphi_1 + \varphi_2)\cos\alpha_0 \cos\vartheta_1 - \sin\alpha_0 \cos(\varphi_1 + \varphi_2); \end{aligned} \tag{46}$$

$$m_{1z}(\vartheta_1) = \cos\alpha_0 \cos\vartheta_1.$$

where: $A = \cos\vartheta_1\left(u_1 \cdot \sin\alpha_0 - r_{g1}\right) + r_{g1}$; $B = u_1 \cdot \cos\alpha_0 + R_{2w}$; u_1, ϑ_1 are linear and angular parameters of the pinion tooth flank; $\varphi_1 = i^* \cdot \varphi_2 = (z_2^* \cdot \varphi_2) / z_1^*$ is the angle of rotation of the pinion (14) during forming the arc tooth flank on the base of the generating wheel.

The movable links of the gear – the pinion and wheel – rotate around axes z_1, and z_2. The pinion and wheel are associated with coordinate systems $S_1(x_1, y_{1,} z_1)$ and $S_2(x_2, y_2 z_2)$. We assume that the starting point of rotation angle ψ_k of the k^{th} movable link (k =1,2) in operating meshing corresponds to the position of axis y_k, (k =1,2) in the axial plane of the gear. The relative position of the pinion and wheel in working meshing (in the absence of rotation) is set by center distance a_{wp}, which differs from the machine distance (a_{ws}) by the value of $\pm\delta a_w$ and teeth twist angle γ.

To study semi-rolled cylindrical gear arc teeth meshing, we determine the position of coordinate system $S_2(x_2, y_2 z_2)$ relative to system $S_1(x_1, y_{1,} z_1)$ using the fourth-order transition matrix $\tilde{D}_{1,2}(\psi_1, \psi_2) = \left\| d_{i,j} \right\|$, $i, j = \overline{1,4}$. The elements of this matrix have the form:

$$d_{11}=\cos\psi_1\cos\psi_2-\sin\psi_1\cos\gamma\sin\psi_2;\quad d_{12}=\cos\psi_1\sin\psi_2+\sin\psi_1\cos\gamma\cos\psi_2;$$
$$d_{13}=\sin\psi_1\sin\gamma;\quad d_{14}=-a_{wp}\sin\psi_1;\quad d_{21}=-\sin\psi_1\cos\psi_2-\cos\psi_1\cos\gamma\sin\psi_2;$$
$$d_{22}=-\sin\psi_1\sin\psi_2+\cos\psi_1\cos\gamma\cos\psi_2;\quad d_{23}=\cos\psi_1\sin\gamma; \qquad (47)$$
$$d_{24}=-a_{wp}\cos\psi_1;\quad d_{31}=\sin\gamma\sin\psi_2;\quad d_{32}=-\sin\gamma\cos\psi_2;\quad d_{33}=\cos\gamma;$$
$$d_{34}=0;\quad d_{41}=d_{42}=d_{43}=0;\quad d_{44}=1.$$

With $\gamma=\delta a_w=0$, the elements of matrix $\tilde{D}_{1,2}(\psi_1,\psi_2)$ (47) when ψ_1 is replaced by φ_1 and ψ_2 is replaced by φ_2 coincide with the elements of matrix $\tilde{A}_{1,2}(\varphi_1,\varphi_2)$ (15). If the function of changing positions of the gear parts

$$\psi_2=\psi_2(\psi_1) \qquad (48)$$

at given values γ and a_{wp} is known, then matrix $\tilde{D}_{1,2}(\psi_1,\psi_2)$ describes the relative movement of the wheel and pinion during the gear operation. In the nonenveloping gear, the law of parameter ψ_2 variation is established after determining the contact points of the active flanks of the pinion and wheel teeth within the single-contact mesh. According to studies [7, 9, 14] of the meshing theory, the contact point on the active flank of the pinion tooth for a fixed value of its rotation angle ($\psi_1=const$) is determined by solving the inverse meshing problem [7, 9, 14], the mathematical description of which is the following system of equations:

$$\tilde{r}_1^1(u_1,\vartheta_1,\varphi_1)=\tilde{D}_{1,2}(\psi_1,\psi_2)\tilde{r}_2^2(u_2,\vartheta_2);$$
$$\tilde{m}_1^1(u_1,\vartheta_1,\varphi_1)=\tilde{D}_{1,2}(\psi_1,\psi_2)\tilde{m}_2^2(u_2,\vartheta_2);\qquad f(u_1,\vartheta_1,\varphi_1)=0. \qquad (49)$$

Here, the superscript defines the coordinate system in which the vector projections are calculated; $\tilde{r}_2^2$, $\tilde{m}_2^2$ are columns matrixes made up of the coordinate projections of radius-vector $\bar{r}_2^2$ (42) and normal unitary vector $\bar{m}_2^2$ (43) of the wheel tooth active flank in coordinate system S_2; $\tilde{r}_1^1$, $\tilde{m}_1^1$ are columns matrixes made up of the coordinate projections of radius-vector $\bar{r}_1^1$ (44) and normal unitary vector $\bar{m}_1^1$ (46) of the wheel tooth active flank in coordinate system S_1; $f(u_1,\vartheta_1,\varphi_1)=0$ is the equation of meshing in processing the pinion teeth flanks (45).

The system (49) corresponds to the conditions of the correct contact of the pinion and wheel teeth active flanks and is equivalent to six scalar transcendental equations (the equality of normal unitary vectors only gives two independent equations) with seven unknowns u_1, ϑ_1, φ_1, u_2, ϑ_2, ψ_1, ψ_2.

$$x_1(u_1,\upsilon_1,\varphi_1) = d_{11}(\psi_1,\psi_2)x_2(u_2,\upsilon_2) + d_{12}(\psi_1,\psi_2)y_2(u_2) + d_{13}(\psi_1)z_2(u_2,\vartheta_2) + d_{14}(\psi_1);$$
$$y_1(u_1,\vartheta_1,\varphi_1) = d_{12}(\psi_1,\psi_2)x_2(u_2,\upsilon_2) + d_{22}(\psi_1,\psi_2)y_2(u_2) + d_{23}(\psi_1)z_2(u_2,\vartheta_2) + d_{24}(\psi_1);$$
$$z_1(u_1,\upsilon_1) = d_{31}(\psi_2)x_2(u_2,\upsilon_2) + d_{32}(\psi_2)y_2(u_2) + d_{33}(\psi_2)z_2(u_2,\vartheta_2) + d_{34}; \qquad (50)$$
$$m_{1x}(\upsilon_1,\varphi_1) = d_{11}(\psi_1,\psi_2)m_{2x}(\upsilon_2) + d_{12}(\psi_1,\psi_2)m_{2y} + d_{13}(\psi_1)m_{2z}(\vartheta_2);$$
$$m_{1z}(\upsilon_1) = d_{31}(\psi_2)m_{2x}(\upsilon_2) + d_{32}(\psi_2)m_{2y} + d_{33}m_{2z}(\vartheta_2); \quad f(u_1,\vartheta_1,\varphi_1) = 0.$$

To determine the coordinates of the contact point of the pinion and wheel teeth flanks with specified errors of the relative position ($\delta a_w, \gamma$) of the pinion and the wheel, it is sufficient to fix the meshing phase ($\psi_1 = const$) within the pinion tooth spacing angle and solve a system of six transcendental equations (50) relative to the unknowns u_1, ϑ_1, φ_1, u_2, ϑ_2, ψ_2. Taking into account that meshing equation $f(u_1,\vartheta_1,\varphi_1) = 0$ is solved analytically (22) with respect to parameter u_1, five transcendental equations remain in the system (50).

The solution of the system (50) is performed numerically using the program developed in MathCad with $\psi_1 = \psi_1^* = const$. Finally, the values of parameters: u_1^*, ϑ_1^*, φ_1^*, u_2^*, ϑ_2^*, ψ_2^* are determined, knowing which allows calculating the projection of the contact point on the wheel and pinion arc teeth flanks using formulas (42) and (44).

To calculate the loading of the contact in meshing arc teeth, it is necessary not only to have the coordinates of the contact point of their flanks, but also the values of the principal relative curvatures at this point. These curvatures characterize the size of the gap between the contacting flanks of the arc teeth up to the value of the second order of smallness in the differential neighborhood of the contact point.

The principal relative curvature in the longitudinal direction of the tooth (k_{p1}) based on the dependences (11) and (39) is calculated as follows:

$$k_{p1} = k_{21} - k_{11}, \qquad (51)$$

where: $k_{21} = -\cos\alpha_0 / \left(u_2^* \cdot \sin\alpha_0 - r_{g2}\right)$.

To determine value k_{11} by the expression (39), we use the formulas (37) and (38), in which we adopt $u = u_1^*$, $\upsilon = \upsilon_1^*$ and $\varphi_2 = \varphi_2^*$.

The principal relative curvature in the profile direction of the tooth (k_{p2}) is the difference between $k_{22} = 0$ and k_{12}:

$$k_{p2} = k_{22} - k_{12} = -k_{12} \qquad (52)$$

where, by analogy with the expression (51), when calculating by the formulas (37), (38) and (39), we adopt values k_{12} as follows: $u = u_1^*$, $\upsilon = \upsilon_1^*$, and $\varphi_2 = \varphi_2^*$.

5 Study of Geometric Characteristics of Semi-rolled Cylindrical Gear arc Teeth Meshing

Based on the built mathematical models for calculating geometric characteristics of meshing semi-rolled cylindrical gear wheel and gear arc teeth, a program is developed in the MathCad software environment. We regard the results of the analysis of the active path of action position in meshing using this program on the example of the study of a semi-rolled cylindrical arc gear, which has the following parameters: $z_1^* = 23$; $z_2^* = 73$; normal modulus $m_n = 10$ mm; tool displacement coefficients when cutting the teeth of the pinion $\chi_1 = 0{,}44$ and the wheel $\chi_2 = 0{,}042$; tooth width $b_w = 120$ mm; $\alpha_0 = 20^o$, radius of the pitch circle of the pinion $R_{w1} = 116{,}115$ mm and the wheel $R_{w2} = 368{,}540$ mm, center distance $a_{wp} = a_{ws} = 484{,}655$ mm. All the calculations are performed for two variants of contact localization in the longitudinal direction of the gear arc teeth. In the first variant (high localization), to cut the concave side of the arc pinion teeth and the convex side of the arc wheel teeth, circular cutting heads with calculated radii of $r_{g1} = 220$ mm and $r_{g2} = 215$ mm respectively are used. In the second variant (the contact is close to linear) they are $r_{g1} = 220$ mm and $r_{g2} = 218$ mm respectively.

The calculation of the points of the active path of action in meshing arc teeth and the principal relative curvatures in the longitudinal direction of the teeth at these points allows estimating the position and size of the contact pattern reflecting the contact point in the gear when generating the gear on the test machine [7]. Following the work [7], the contact pattern corresponds to the line of the level of the gap occurring in the vicinity of the contact point of the flanks and calculated by the formula:

$$\Delta = 0{,}006\sqrt{m_n}\,. \tag{53}$$

Value Δ is associated with k_{p1} by the relationship:

$$\Delta = k_{p1} \cdot z_b^2 / 2, \tag{54}$$

where: z_b is the half-width of the contact area.

Combining (53) and (54), we obtain:

$$z_b = \sqrt{2 \cdot \Delta / k_{p1}} = \sqrt{(0.012\sqrt{m_n}) / k_{p1}} \,. \tag{55}$$

Expression (55) allows calculating the half-length of the contact area in each phase of meshing the arc teeth and determining the size of the contact pattern relative to the contact point.

Figures 5-6 show the results of calculating the contact pattern for gear 1 and gear 2 in the presence of the teeth twist angle $\gamma = 0.0015$.

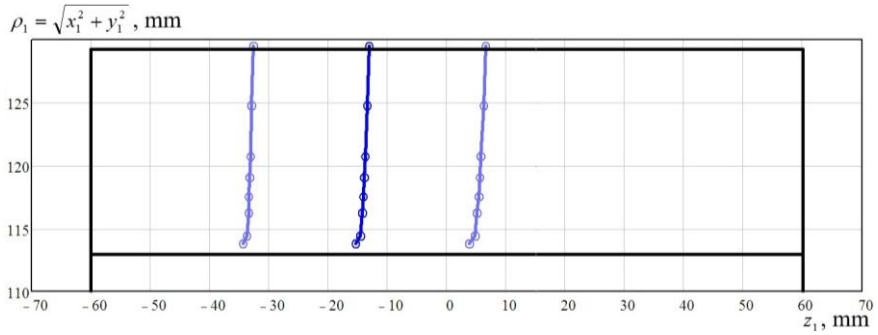


Figure 5
Active path of action and contact pattern in a cylindrical arc teeth gear (variant 1) at γ =0.0015

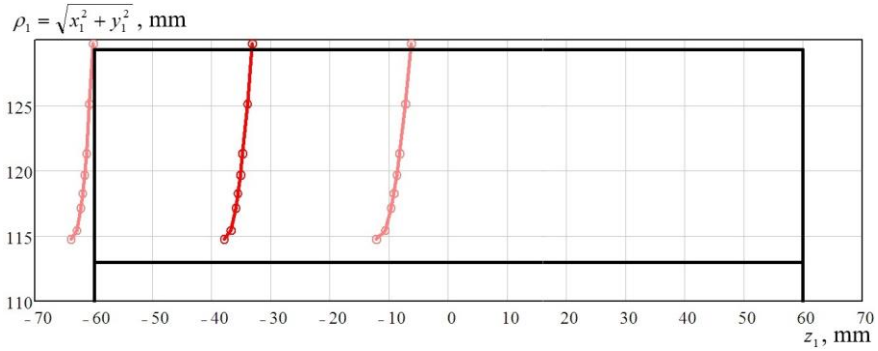


Figure 6
Active path of action and contact pattern in a cylindrical arc teeth gear (variant 2) at γ =0.0015

In the gear manufactured according to variant 2, the contact surface integrity of the pinion and wheel teeth is higher than in variant 1. As a result, the width of the contact pattern has increased and is $\approx 53\%$ in the absence of misalignment. At the same time, the sensitivity of the gear to the tooth flanks misalignment has increased. The left border of the contact pattern at a twist angle of γ =0.0015 already "goes" to the edge of the tooth (Figure 6), while the width of the contact pattern decreases by $\approx 20\%$, which will cause an increase in contact stresses.

At a twist angle of γ =0.0021, despite the presence of contact between the tooth surfaces, the width of the contact pattern is sharply reduced (by almost 50%). The gear will operate essentially under edge contact conditions.

Conclusions

In this work, mathematical models of forming the gear wheel and pinion arc teeth flanks have been built. Dependences for calculating principal curvatures of the wheel and pinion arc teeth flanks have been obtained. An algorithm and a program for calculating the coordinates of the active path of action points in meshing arc teeth as well as principal relative curvatures at these points have been developed. The problem of determining the size of the contact pattern in the studied gear has been solved.

The calculated geometric characteristics of meshing gear arc teeth in the presence of a twist angle of the wheel and pinion teeth in meshing are the basis for calculating the loading capacity of the gear, which depends both on the position of the contact points of the teeth in meshing, principal relative curvatures at these points, and the maximum value of the twist angle of the teeth in meshing, when the performance and durability of the gear are to be provided.

List of notations

$S_p(x_p, y_{p,} z_p)$ - coordinate system rigidly connected to the cutting head;

u, ϑ - linear and angular parameters of the generating surface;

α_0 - basic profile angle;

r_{g2} - calculated radius of the cutting head at wheel cutting;

$S_2(x_2, y_2 z_2)$ - coordinate system rigidly connected to the wheel;

$\tilde{A}_{2,p}$ - matrix describing the transition from coordinate system $S_p(x_p, y_{p,} z_p)$ to system $S_2(x_2, y_2 z_2)$;

$\bar{r}_p(x_p, y_{p,} z_p)$ - radius-vector of a generating surface;

$\bar{r}_2(x_2, y_{2,} z_2)$, $\overline{m}_2(m_{2x}, m_{2y}, m_{2z})$ - radius vector and normal unitary vector of the arc wheel tooth flank;

$\tilde{r}_2$, $\tilde{r}_p$ - columns matrixes made up of coordinates of $\bar{r}_2(x_2, y_{2,} z_2)$ and $\bar{r}_p(x_p, y_{p,} z_p)$;

R_{w2} - radius of the wheel pitch circle;

m_{2x}, m_{2y}, m_{2z} - projections of the normal unitary vector of arc wheel tooth flank;

$d\overline{m}_2$, $d\bar{r}_2$ -differentials of vectors $\overline{m}_2$ and $\bar{r}_2$;

x_{2u}; z_{2u}; $x_{2\vartheta}$; $z_{2\vartheta}$; m_{2xu}; $m_{2x\vartheta}$; m_{2zu}; $m_{2z\vartheta}$ - partial derivatives of projections of coordinate of vectors $\bar{r}_2$ and $\overline{m}_2$ by parameters u and ϑ;

k_{21}, k_{22} - the principal curvatures of the arc wheel flank;

z_1^*, z_2^* - the numbers of the pinion and wheel teeth;

$S_1(x_1, y_1 z_1)$-coordinate system rigidly connected to the pinion;

φ_2 - the angle of the wheel rotation when forming the pinion tooth flank;

$\tilde{A}_{1,2}(\varphi_2)$ - matrix describing the transition from coordinate system $S_2(x_2, y_2 z_2)$ to coordinate system $S_1(x_1, y_1 z_1)$;

a_{ws} - the center distance in the machine meshing of the pinion and the generating wheel;

$\bar{r}_1(x_1, y_{1,} z_1)$, $\overline{m}_1(m_{1x}, m_{1y}, m_{1z})$ - radius vector and normal unitary vector of the arc pinion tooth flank;

$f(u, \vartheta, \varphi_2) = 0$ the meshing equation;

$\overline{V}_\varphi$ - the vector analog of the relative speed;

$\tilde{C}_\varphi$ - the fourth-order matrix of the relative speed analog;

$d\overline{m}_1$, $d\bar{r}_1$ - differentials of vectors $\overline{m}_1$ and $\bar{r}_1$;

f_u, f_ϑ, f_φ - partial derivatives of the meshing $f(u, \vartheta, \varphi_2) = 0$ by parameters u, ϑ and φ_2;

k_{11}, k_{12} - principal curvatures of the pinion tooth flank;

ψ_1, ψ_2 the angles of rotation of the pinion and wheel in working meshing;

γ - twist angle of teeth in meshing;

k_{p1}, k_{p2} - principal relative curvatures in arc teeth meshing ;

z_b - semi-length of contact pattern;

r_{g1} - calculated radius of the cutting head at pinion cutting.

References

[1] Syzrantsev V. N.: Cylindrical Arc Gears: History, Achievements, and Problems, *Gears in Design, Production and Education. Mechanisms and Machine Science* 101 (2021) 131-151, https://doi.org/10.1007/978-3-030-73022-2_6

[2] Syzrantsev V., Syzrantseva K., Varshavsky M.: Contact Load and Endurance of Cylindrical Gearing with Arch-shaped Teeth, *International Conference on Mechanical Transmissions (ICMT 2001)*, Chongqing, China (2001) 425-431

[3] Arafa, H. A.: C-gears: Geometry and Machining, *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219, Issue 7 (2005) 709-726

[4] Tsay G.-B., Fong Z. H.: Tooth Contact Analysis for Helical Gears with Pinion Circular Arc Teeth and Gear Involute Shaped Teeth, *Journal of Mechanical Design, Transactions of the ASME* 111, Issue 2 (1989) 278-284

[5] Tseng J.-T., Tsay C.-B.: Mathematical Model and Surface Deviation of Cylindrical Gears with Curvilinear Shaped Teeth Cut by Hob Cutter, *Journal of Mechanical Design, Transactions of the ASME* 127, Issue 5 (2005), 982-987

[6] Dai Y., Yukinori A., Jiang D.: Hobbing Mechanism of Cylindrical Gear with Arcuate Tooth Traces and Experimental Investigation, *Zhongguo Jixie Gongcheng / China Mechanical Engineering,* 17, Issue 7 (2006) 706-709

[7] Litvin F. L., Fuentes A.: Gear Geometry and Applied Theory, Cambridge: University Press, New York (2004) 800 p.

[8] Stadtfeld H. J.: Gleason Bevel Gear Technology. Manufacturing, Inspections and Optimization, The Gleason Works, Rochester, New York (1995) 202 p.

[9] Litvin F. L., Gutman, Y.: Methods of Synthesis and Analysis for Hypoid Gear Drives of Formate and Helixform. Parts 1, 2, and 3, *ASME Journal of Mechanical Design* 103, No. 1 (1981) 83-113

[10] Fan Q.: Computerized Modeling and Simulation of Spiral Bevel and Hypoid Gears Manufactured by Gleason Face Hobbing Process, *Journal of Mechanical Design, Transactions of the ASME* 128 (6) (2006) 1315-1327

[11] Syzrantseva K., Syzrantsev V., Babichev D.: Comparative Analysis of Stress-Strain Condition of Cylindrical Gears Arc Teeth and Spurs, *Lecture Notes in Mechanical Engineering* (2020) 101-108

[12] Syzrantseva K. V., Syzrantsev V. N., Kolbasin D. S.: Comparative Estimation of the Failure Probability of Cylindrical Arc and Helical Gears by Tooth Bending Endurance, *AIP Conference Proceedings* (2019) 2176:020010, doi.org/10.1063/1.5135122

[13] Syzrantsev V., Syzrantseva K., Kolbasin D.: Forming Surfaces of a Semi-rolled Cylindrical Gearing Wheel and a Gear Arc Teeth, *Lecture Notes in Mechanical Engineering* (2021) 134-141, https://doi.org/10.1007/978-3-030-54814-8_16

[14] Syzrantsev Vladimir, Syzrantseva Ksenia: Study of Geometric Characteristics of the Arc Teeth Semi-rolled Cylindrical Gear Meshing, *FME Transactions* 49 (2021) 367-373 doi:10.5937/fme2102367S

[15] Bálint Laczik, Peter Zentay, Richárd Horváth: A New Approach for Designing Gear Profiles using Closed Complex Equations, *Acta Polytechnica Hungarica* 11 (2014) 159-172

New Numerical Procedure for Determination of Elastic Curve of Statically Determinate and Indeterminate Beams with Variable Cross Sections

István Bíró

University of Szeged, Faculty of Engineering, Mars tér 7, H-6724 Szeged, Hungary, biro-i@mk.u-szeged.hu

Abstract: In this paper a new numerical procedure is developed for calculating the inclination angle and deflection as initial conditions of the end points of statically determinate and indeterminate beams. The method is based on the topology comparison of simple (hinge-roller combination) supported beam and a resemblant cantilever beam. Assuming that the support reactions of the beam are active forces, the virtual displacements at the points of the reaction forces are calculated. Based on these values the inclination angle is calculated. Several examples are considered and the suggested in this paper, while the procedure is applied for various types of structures and loadings. The results, obtained by the suggested numerical procedure, are compared with analytical ones, and they are in good agreement.

Keywords: elastic curve; beams of variable cross section; initial guess for slope and deflection

1 Introduction

Beam-like members [1] such as shafts, levers, frame components, beam structures, etc. are regularly designed and constructed in the field of mechanical and civil engineering. Before construction or fabrication of the structure the knowledge of the integrity, i.e., the deflection and inclination angle of the structure, is pertinent. The inclination angle has also the importance in detection of the modal parameters of the beam which seem to be of great significance in bridge and other structural health detection and damage identification [2]. For various values of inclination angles the failure modes of components are computed and the evaluation of failure with increasing the angle is studied. Inclination angles are obtained applying the theoretical approach or using experimental methods. For example the inclination angle is measured by Yang and Qin [3] with the inclinometer. However, the experimental procedure is complex and connected with troubles and costs.

To overcome these problems numerous methods for determination of the inclination angle are developed [4-10]. Thus, the large deflection of a simply supported beam loaded in the middle has been studied analytically by using the exact solutions and the finite element method. In practice, the inclination angles are computed applying the commercial simulation packages. Recently, a computational tool, CABDA, has been designed and developed on MATLAB where the algorithm is based on analytic equations of beam deflection [1]. The program is tested on steel and brass rectangular beams and the results are compared with those obtained experimentally and by simulation. Some differences in the results have been observed. The error in numerically obtained solutions is explained with the fact that the program uses the linear structure theory, which is not applicable for strongly nonlinear systems. If the deformation of the beam is small, the use of linear theory for determining the shape of the elastic curve and the inclination angle is appropriate. However, the results obtained according to the linear theory are not convenient for the beam with large deformation and strong nonlinearity. In these special cases, modification in the numerical solving procedure is necessary and the nonlinear structure theory has to be included.

Recently, some analytical investigations on calculation of inclination angle of strong nonlinear structures were carried out and published. Thus, the inclination angle of a prismatic cantilever beam subjected to a combination of inclined end force and tip moment was computed by Abu-Alshaikh et al. [11]. The nonlinear theory of bending and the exact expression of the curvature are used. Based on an elliptic integral formulation, an accurate numerical solution is obtained. Comparing with previously published results, the accuracy of numerical solution obtained with the method is more accurate. In terms of Jacobi elliptic functions, the solution of equilibrium configuration of an elastic beam, subjected to three-point bending, is given by Batista [12]. Results obtained numerically are compared with those of other authors. The relationship between force and deflection of a thin elastic beam is given approximately as a polynomial function. The Galerkin method is used to obtain an approximate force-deflection characteristic of the [13].To validate the result the exact solution and that from the finite element method are used. The analytic Homotopy Perturbation Method (HPM) is adopted by Hatami et al. [14] for predicting the deflection of a cantilever beam subjected to static co-planar loading. The analytical solution procedure is applied for a Reissner's beam underforce acting at the midpoint be-tween two supports [15]. Comparing HPM through numerical results it is demonstrated that HPM can be a high efficiency procedure for computing the deflection. However, the procedure is rather complex and the computation requires significant time. Machado et al. [16] introduced a weighted algorithm, based on the reduced differential transform method. The proposed scheme considers the initial and boundary conditions simultaneously for obtaining a solution of the equation.

To overcome the computation problem, the aim of this paper is to introduce a new procedure for calculating the inclination angle and deflection as initial conditions of the end points of statically determinate and indeterminate beams. The numerical procedure would involve less computational time compared with other techniques available in literature. The method is based on a topology comparison of a simply supported beam and its resemblant or to say "modified" cantilever beam. Assuming that the support reactions of the beam are active forces, the virtual displacements at the points of the reaction forces are calculated. Based on these values the inclination angle and deflection of the endpoint of the beam as initial conditions can be calculated. Several examples are considered and the suggested in this paper, while the procedure is applied for various types of structures and loadings. The results, obtained by the suggested numerical procedure, are compared with analytical ones, and they are in good agreement.

As problem solving technique this numerical method can be offered for engineers. It can be treated as one of problem solving techniques for engineers published by Horvath and Rudas [17]. The demonstrated method can be applied to metamaterial beams as well published by Cveticanin and Mester [18].

The paper is divided into five sections. After the introduction, in Section 2, the theorem of calculation of the inclination angle at the end point of the beam is introduced and proved. In Section 3, the procedure of transforming the boundary value problem into initial problem for a simply supported to a cantilever beam with variable cross-section is presented. The procedure is applied on examples. The paper ends with conclusions.

2 Procedure for Computing of the Approximate Inclination Angle

Theorem 1. In case of linear model of simple supported beams with two consoles loaded at arbitrary places by concentrated and/or distributed forces and/or couple of forces the inclination angle of free end of the console on the left side is

$$\varphi_C = -\frac{y_B - y_A}{l}, \tag{1}$$

where l is the distance between the supports, y_A and y_B are the elastic deflections at cross-section A and B of the "modified" beam. The "modified" beam is clamped at cross-section C, with an identic active load system compared to the original model.

It must be mentioned that the calculated reaction forces are considered active forces in the modified version of the model. Applied notations can be seen in Figure 1.

There are 6 different loading components of the beam in Figure 1 such as concentrated forces and couple of forces acting on consoles on the right or left side or between the supports.

To proof of the theorem/equation (1) is investigated with regard to the different load-ings of the consoles, the effective span of beam together with the uniform and variable cross-section. Theorem/equation (1) is proved for each load cases. Based on the superposition, the principle of the theorem/equation (1) is true for any simple supported beams loaded by concentrated and couple of forces at any places.

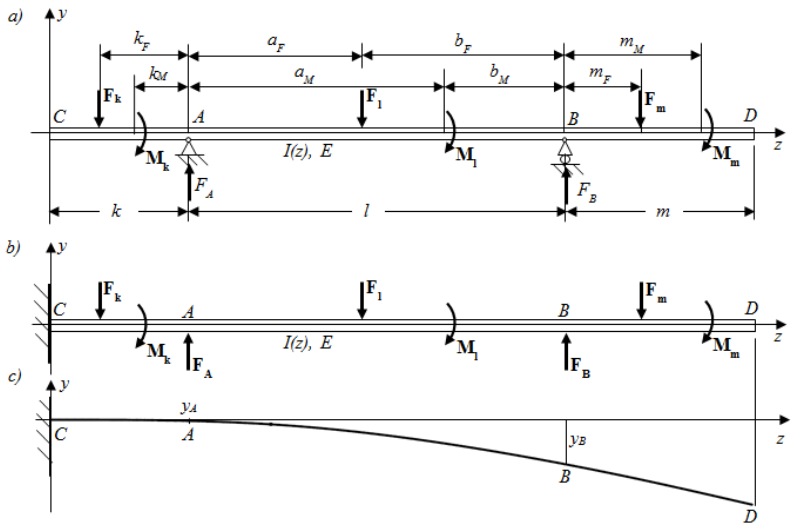


Figure 1

(a) Scheme of simple supported beam with two consoles loaded at arbitrary places by concentrated forces and couple of forces (notations to the proof); (b) „Modified" beam clamped at cross section C; (c) Shape of the elastic curve in case of modified beam (φ_{Cmod}=0, y_{Cmod}=0)

Proof of Theorem 1. The proof of the theorem (1) is done on examples of beams with various types of loadings shown in Figures 1, 12-17. The formula for calculating the elastic curves, y(z), of beams of variable cross-section and loaded with bending moment M(z) is

$$EI(z)\frac{y''(z)}{(1+y'^2(z))^{\frac{3}{2}}} = -M(z), \tag{2a}$$

where EI(z) is flexural rigidity of the beam, E is modulus of elasticity, I(z) is moment of in-ertia of the cross section about its neutral axis, M(z) represents the bending moment func-tion of the beam, z is the position coordinate, while (‘)=d/dz and (“)= d^2/dz^2. In our calculation the linearized version is applied

$$EI(z)y''(z) = -M(z). \tag{2b}$$

Based on the principle of superposition, the theorem (1) has to be proved. In Figure 1a), a three-part beam, i.e. two consoles and an effective span is shown.

The simple supported beam with two consoles is loaded at arbitrary places by concentrated forces and couple of forces. In Figure 1b), the „modified" beam clamped at cross section C can be seen. In Figure 1c), the shape of the elastic curve in case of modified beam ($\varphi_{Cmod}=0$, $y_{Cmod}=0$) is shown.

In Table 1 the results for different types of supported beams loaded with various types of loading are presented. The applied notations in the Table 1 and Figures 12-17 are:

- F_A and F_B are reaction forces in cases of active loading,
- y_A and y_B are the elastic deflections of cross section A and B in case of the „modified" beam,
- and φ_C is the inclination angle calculated according to equation (1).

Table 1

Summary of physical quantities to prove the presented theorem in case of beam of uniform cross-section

Model	F_A	F_B	y_A
Figure 12	$\frac{F_k(k_F+l)}{l}$	$\frac{F_k k_F}{l}$	$-\frac{F_k k_F^3}{6IE}$
Figure 13	$\frac{M_k}{l}$	$\frac{M_k}{l}$	$-\frac{M_k k_M^2}{2IE}$
Figure 14	$\frac{F_l b_F}{l}$	$\frac{F_l a_F}{l}$	0
Figure 15	$\frac{M_l}{l}$	$\frac{M_l}{l}$	0
Figure 16	$\frac{F_m m_F}{l}$	$\frac{F_m(m_F+l)}{l}$	0
Figure 17	$\frac{M_m}{l}$	$\frac{M_m}{l}$	0

Model	y_B	φ_C
Figure 12	$-\frac{F_k(k_F+l)k_F(k_F+2l)}{6IE}$	$-\frac{F_k k_F(3k_F+2l)}{6IE}$
Figure 13	$-\frac{M_k(3k_M^2+6k_M l+2l^2)}{6IE}$	$\frac{M_k(l+3k_M)}{3IE}$
Figure 14	$-\frac{F_l a_F b_F(a_F+2b_F)}{6IE}$	$\frac{F_l a_F b_F(a_F+2b_F)}{6lIE}$
Figure 15	$-\frac{M_l(2b_M^2-2a_M b_M-a_M^2)}{6IE}$	$\frac{M_l(2b_M^2-2a_M b_M-a_M^2)}{6lIE}$
Figure 16	$\frac{F_m m_F l^2}{6IE}$	$-\frac{F_m m_F l}{6IE}$
Figure 17	$\frac{M_m l^2}{6IE}$	$-\frac{M_m l}{6IE}$

Important remark: φ_C inclination angle of the real beam is determined directly by applying the Betti-theorem. The inclination angle in every single case is presented in the last column (Table 1).

In previous load cases it can be seen that the inclination angle of cross-section φ_C can be determined with the arbitrary lengths of the consoles and the effective span. Moreover, φ_C is independent from the positions of the different loadings. In the above-mentioned load cases the flexural rigidity (IE) of the beam is constant along axis z.

It must be mentioned that equation (1) is valid for beams of variable cross-section as well. As it can be seen in Figure 12-17, concentrated and couple of forces act on the left or right consoles or between the supports.

According to this fact, concerning beams of arbitrary variable cross-section, there are different load cases demonstrated in Figure 18. (See in Appendix.) Different types of statically determinate beams of variable cross-sections, with various types of loading, are considered, while the results are presented in Table 2.

Table 2

Summary of physical quantities to prove the presented theorem in case of beams of variable cross-section

Model	$\mathbf{F_A}$	$\mathbf{F_B}$	$\mathbf{y_A}$
Figure 18 (a)	$\frac{F(k+l)}{l}$	$\frac{Fk}{l}$	$-\frac{F}{E}\int_0^k \frac{kz-z^2}{I_k(z)}dz$
Figure 18(b)	$\frac{M}{l}$	$\frac{M}{l}$	$\frac{M}{E}\int_0^k \frac{k-z}{I_k(z)}dz$
Figure 18(c)	$\frac{Fb}{a+b}$	$\frac{Fa}{a+b}$	0
Figure 18(d)	$\frac{M}{a+b}$	$\frac{M}{a+b}$	0

Model	$\mathbf{y_B}$
Figure 18(a)	$-\frac{F}{E}\int_0^k \frac{(k+l)-z^2}{I_k(z)}dz - \frac{Fk}{lE}\int_0^l \frac{z^2}{I_l(z)}dz$
Figure 18(b)	$\frac{M}{E}\int_0^k \frac{k+l-z}{I_k(z)}dz + \frac{M}{lE}\int_0^l \frac{z^2}{I_l(z)}dz$
Figure 18(c)	$\frac{Fa}{(a+b)E}\int_0^b \frac{z^2}{I(z)}dz - \frac{Fb}{(a+b)E}\int_b^{a+b} \frac{z^2-(a+b)z}{I(z)}dz$
Figure 18(d)	$-\frac{M}{(a+b)E}\int_0^a \frac{(a+b)-z^2}{I(z)}dz + \frac{Fb}{(a+b)E}\int_a^{a+b} \frac{((a+b)-z)^2}{I(z)}dz$

Model	φ_C
Figure 18(a)	$\frac{F}{E}\int_0^k \frac{z}{I_k(z)}dz + \frac{Fk}{l^2E}\int_0^l \frac{z^2}{I_l(z)}dz$
Figure 18(b)	$-\frac{M}{E}\int_0^k \frac{1}{I_k(z)}dz - \frac{M}{l^2E}\int_0^l \frac{z^2}{I_l(z)}dz$
Figure 18(c)	$-\frac{Fa}{(a+b)^2E}\int_0^b \frac{z^2}{I(z)}dz + \frac{Fb}{(a+b)^2E}\int_b^{a+b} \frac{z^2-(a+b)z}{I(z)}dz$
Figure 18(d)	$\frac{M}{(a+b)^2E}\int_0^a \frac{(a+b)-z^2}{I(z)}dz - \frac{Fb}{(a+b)^2E}\int_a^{a+b} \frac{((a+b)-z)^2}{I(z)}dz$

Applied notations in the head of Tables 1 and 2 are the same. The inclination angle φ_C in all of cases is determined in both ways again.

In Figure 18a)-b) the console is subjected to loads on its left side. Due to the symmetry, it is enough to prove equation/theorem (1) for inclination angle φ_B. In this case, the beam is clamped at cross section B.

Applying notations of Figure 18a, inclination angle of cross section B of the original beam,

$$\varphi_B = \frac{Fk}{l^2E}\int_0^l \frac{zl-z^2}{I(z)}dz. \tag{3}$$

In case of the "modified" beam, i.e. our current beam is clamped at cross section B and loaded by concentrated force and reaction forces of simple supported beam (Figure 18a) the elastic deflections at cross section A and B can be determined on the basis of the Betti-theorem,

$$y_A = -\frac{Fk}{lE}\int_0^l \frac{zl-z^2}{I(z)}dz, \quad y_B = 0. \tag{4}$$

According to equation (1),

$$\varphi_B = -\frac{y_A-y_B}{l} = -\frac{1}{l}\left[-\frac{Fk}{lE}\int_0^l \frac{zl-z^2}{I(z)}dz - 0\right] = \frac{Fk}{l^2E}\int_0^l \frac{zl-z^2}{I(z)}dz, \tag{5}$$

which complies with equations (1) and (3).

Applying notations of Figure 18b, inclination angle of cross section B of real beam,

$$\varphi_B = -\frac{M}{l^2E}\int_0^l \frac{zl-z^2}{I(z)}dz. \tag{6}$$

In case of the "modified" beam, i.e. – in this case – clamped at cross section B loaded by couple of forces and reaction forces of simple supported beam (Figure 18b) the elastic deflections at cross-section A and B with regard to the Betti-theorem,

$$y_A = \frac{M}{lE}\int_0^l \frac{zl-z^2}{I(z)}dz, \quad y_B = 0. \tag{7}$$

According to equation (1),

$$\varphi_B = -\frac{y_A - y_B}{l} = -\frac{1}{l}\left[\frac{M}{lE}\int_0^l \frac{zl-z^2}{I(z)}dz - 0\right] = -\frac{M}{l^2E}\int_0^l \frac{zl-z^2}{I(z)}dz \tag{8}$$

which complies with equations (1) and (6).

Comparing the results, obtained by the Betti-theorem and equation (2), the theorem is proved.

3 Method of Transformation Boundary Value Problem into Initial Value Problem

In Chapter 2 the proof of theorem/equation (1) for statically determinate beams with uniform and/or variable cross-sections, loaded by different way, can be seen. φ_C inclination angle is the initial slope of the statically determinate (original) beam. Based on the superposition principle, the effect of active load components of the beam are independent from each other. Therefore, the theorem/equation (1) is valid independently in the linear dimension.

Based on the theorem/equation (1), the elastic curve of statically determinate beams can be determined by the following steps:

- Calculation of reaction forces,
- Determination of moment function M(z) of the beam,
- Substitution of the moment function into differential equation (2b),
- Numerical solution of the differential equation with initial conditions y(0)=0, y'(0)=0. At this step the beam is treated as „modified", i.e. it is clamped at cross section C,
- Applying obtained deflections y_A and y_B the initial slope, $\varphi_C = -(y_B - y_A)/l$,
- Repeating the numerical process with initial conditions y(0)=0, y'(0)=φ_C values of deflections y_A and y_B which are obtained similarly,
- Repeating the numerical process with initial conditions y(0)=-y_A, y'(0)=φ_C. As a result, the shape of the elastic curve of the real beam is obtained.

4 Results

4.1 Simply Supported Beam

In presented numerical examples equation (2a) is applied.

Example 1 As an example, the task is to determine numerically the elastic curve of cantilevered simply supported beam shown in Figure 2.

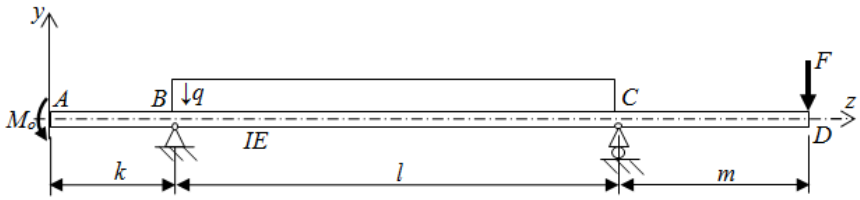


Figure 2
Cantilevered simply supported beam

The following numerical data are given: F=2000 N, q=4000 N/m, M_o=4000 Nm, k=1000 mm, l=3500 mm, m=1500 mm, E=210 GPa, I=328 cm^4.

For the given numerical values the reaction forces of the supports:

$$F_B = \frac{1}{l}\left(M_o + \frac{q}{2}l^2 - Fm\right),$$

$$\mathrm{F_C} = \frac{1}{l}\left(F(l+m) - M_o - \frac{q}{2}(k^2 - (l+m)^2\right), \tag{9}$$

while the moment(z) function is plotted in Figure 3.

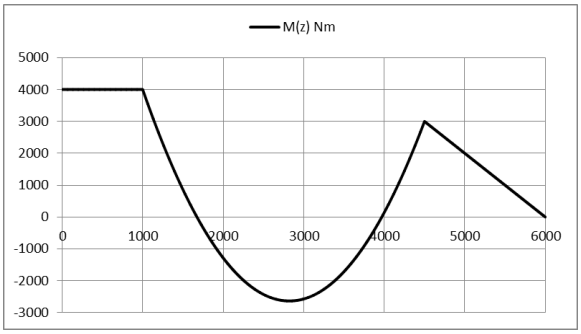


Figure 3
Moment function(z) of the cantilevered simply supported beam

Three segments along the beam are evident and the differential equations according to (2a) of the elastic curve for each segment are formed. The obtained relations are

$$0 \le z \le k, \quad y_1'' = -\frac{M_0}{IE}(1+y'^2)^{\frac{3}{2}} \tag{10}$$

$$k \le z \le k+l,$$

$$y_2'' = -\frac{1}{IE}\left(\frac{q}{2}z^2 - \left(\frac{q}{2}l + \frac{M_0 - Fm}{l} + qk\right)z + M_0\frac{l+k}{l} - \frac{Fmk}{l} + \frac{q}{2}k^2 + \frac{q}{2}lk\right)(1+y'^2)^{\frac{3}{2}}, \tag{11}$$

$$k+l \le z \le k+l+m, \quad y_3'' = -\frac{F}{IE}(k+l+m-z)(1+y'^2)^{\frac{3}{2}}. \tag{12}$$

Let us solve the above-mentioned equations numerically for initial values $y'_o=0$ and $y_o=0$. Namely, it is assumed that the left end of the beam is fixed and corresponds to a cantilever. Therefore, the moment function as a function of coordinate z does not yet correspond with the original beam. The obtained result is plotted in Figure 4.

Obviously the shape of the elastic curve is not suitable to the original loading and the constraint relations. In order to get the accurate initial values let us carry out the following transformations.

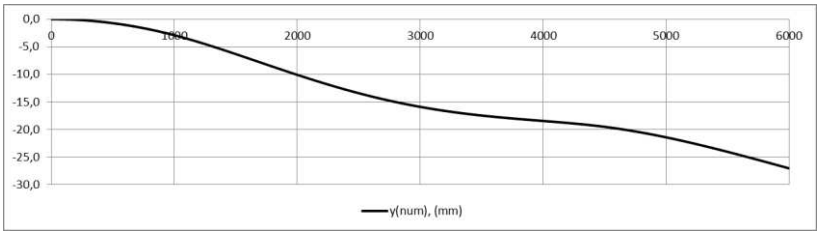


Figure 4

Elastic curve, as a function of z, of the cantilever for initial conditions y'o=0 and yo=0

Rotation around axis perpendicular to xy plane

Creating the ratio of differences between deflections of cross-sections B and C and between their positions coordinates an angle can be obtained as follows:

$$\varphi = \frac{y_C - y_B}{z_C - z_B} = \frac{-19{,}52532706\text{mm} - (-2{,}90391530)\text{mm}}{4500\text{ mm} - 1000\text{ mm}} = -0{,}0047489747\text{ rad.} \tag{13}$$

This angle with opposite sign can be treated as initial inclination of cross-section A, i.e.

$$y'_A = -\varphi = 0{,}0047489747\text{ rad.} \tag{14}$$

The numerical calculation of differential equations of the elastic curves is repeated with initial values:

$$y_A = 0,\ y'_A = 0{,}0047489747 \text{ rad}. \tag{15}$$

The obtained elastic curve is plotted in Figure 5.

It can be noticed that for initial conditions (15) the values of deflection at supports B and C are equal: $y_B = y_C = 1{,}84504132$ mm. After this recognition translation along axis y seems obvious.

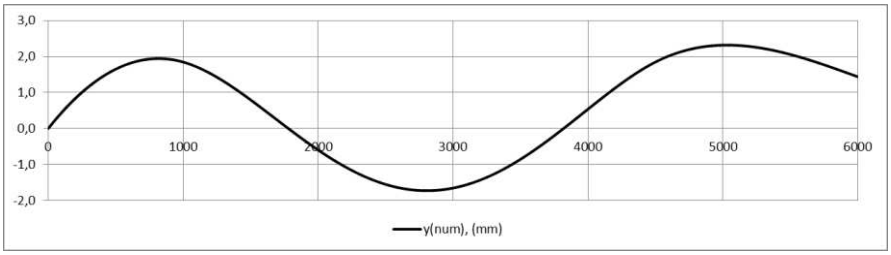


Figure 5

Elastic curve for initial values: $y_A = 0\ mm,\ y'_A = 0{,}0047489747\ rad$

Translation along axis y

Now, the curve is translated along y axis for the value

$y_A = -y_B = -y_C = -1{,}84504132\ mm$, to move the supports in the position with zero deflection. Starting with numerical procedure and applying the calculated initial values $y_A = -1{,}84504132$ mm, $y'_A = 0{,}0047489747$ rad the elastic curve of the beam are obtained and plotted in Figure 6. In order to check the obtained results the Betti-theorem (Table 3) is applied. Results obtained in different ways are compared to each other and summarized in Table 3.

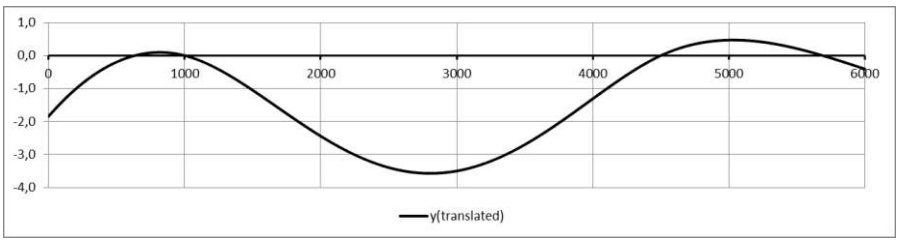


Figure 6

Elastic curve determined based on numerically and analytically obtained initial conditions

4.2 Cantilevered Simply Supported Beam having Sinusoidal Variable Circular Cross-Section

Example 2 The sketch of the cantilevered simply supported beam can be seen in the previous example. In this case, there is a beam having variable circular cross-section (Figure 7). Its diameter is described by equation d(z)=100+30sin(0.004712z) [mm]. Other input data are the same.

Starting with numerical procedure again and applying the calculated initial values $y_A = -2{,}98325081$ mm, $y'_A = 0{,}00433389$ rad the elastic curve of the beam is obtained and plotted in Figure 8. In order to check the obtained results the Betti-theorem is applied

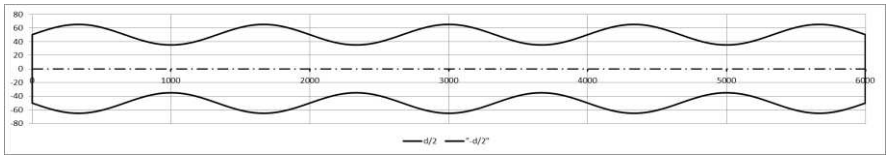


Figure 7
Shape of the simply supported beam having variable circular cross-section (side view)

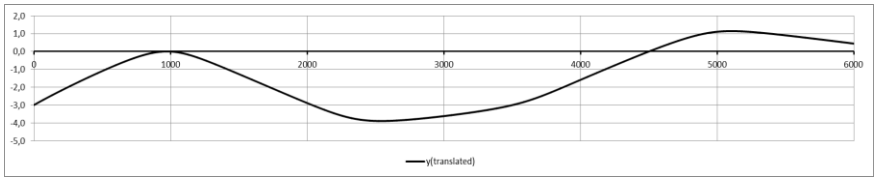


Figure 8
Elastic curve determined and based on obtained initial conditions

Results obtained in different ways are compared to each other and summarized in Table 3. Comparing the results obtained numerically for the nonlinear model and analytically for the linearized system (Betti-theorem) it can be concluded that the difference between them is negligible, moreover it can be seen the effect of nonlinearity is negligible as well.

Table 3
Comparison of deflections of cross-sections A and D obtained in different ways

Example 4.1	**y_A, mm**	**y_D, mm**
Numerical transformation method - rotation and translation - transformation of the boundary value problem into the initial value problem	-1.8452	-0,4083
Betti-theorem for beams of variable cross-section	-1.8398	-0,4127

Example 4.2	y_A, mm	y_D, mm
Numerical transformation method - rotation and translation - transformation of the boundary value problem into the initial value problem	-2,9832	0,4438
Betti-theorem for beams of variable cross-section	-2,9830	0,4408

4.3 Cantilevered Statically Indeterminate Beam to the First-Degree having Variable Circular Cross-Section

Example 3. As third example a statically indeterminate beam with three supports, having variable circular cross-section is shown in Figure 9. The task is the same: to determine numerically the elastic curve of the beam. Following numerical data are given: F_1=6000 N, F_2=16000 N, q=12000 N/m, M_o=3000 Nm, a=1000 mm, E=210 GPa.

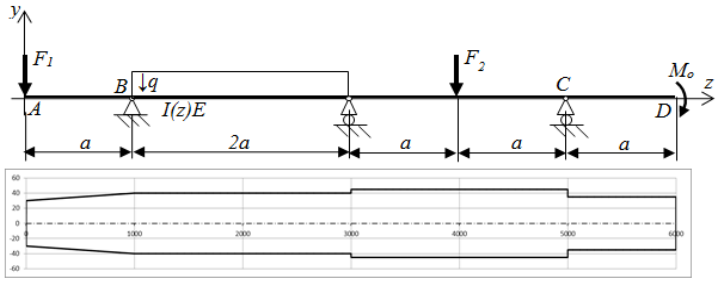


Figure 9

Cantilevered beam with three supports together with the shape of the of the beam having a variable circular cross-section (side view)

As a result of applying the Clapeyron-equation the moment(z) function can be seen in Figure 10.

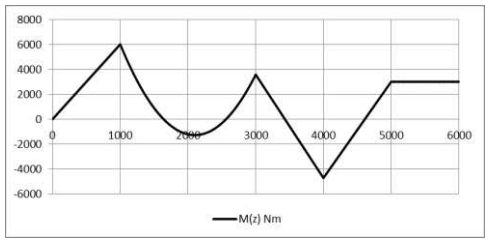


Figure 10

Moment function of the cantilevered beam with three supports

By applying of above described numerical procedure again, calculated initial values are y_A=9,1415991914 mm, y'_A=0,01335125 rad. Obtained elastic curve of the beam plotted in Figure 11.

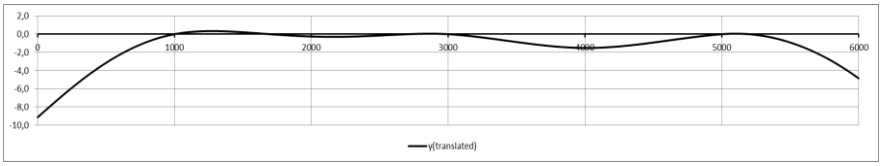


Figure 11
Elastic curve determined and based on obtained initial conditions

Conclusions

It can be concluded:

- The initial slope of the arbitrary loaded simple supported beam can be determined with high accuracy if the structure is modified into a clamped-free beam. For that case the inclination angle of the free end of the beam is the ratio between the difference of elastic deflections of cross sections in the supporting points of the 'modified beam' and the distance between supports.
- For the case of small deformation when the nonlinearity is weak the suggested procedure for calculation of the inclination angle is applicable with certain accuracy.
- However, if the deformation is large and the nonlinearity is strong serious number of iterative steps are necessary to reach the demanded accuracy.
- Applying the suggested formula for inclination angle the elastic curves of simple supported beams can be determined numerically.
- Based on the suggested procedure the boundary value problem of simple supported and continuous beam is transformed into initial value problem which is a special and effective application of the shooting method. The method is stable and easy to use.
- Results obtained by the method and compared with those obtained with Betti theorem for the linear models show a good agreement.

Appendix A

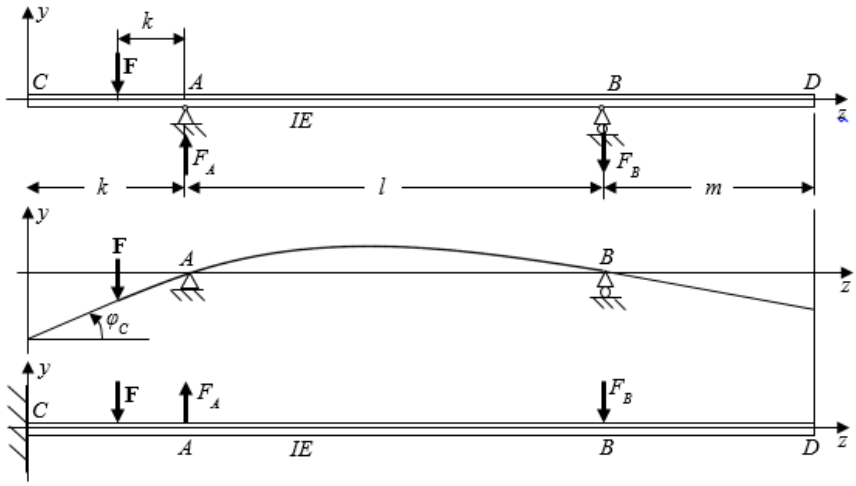


Figure 12

Simple supported beam loaded by concentrated force on the left console at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section C

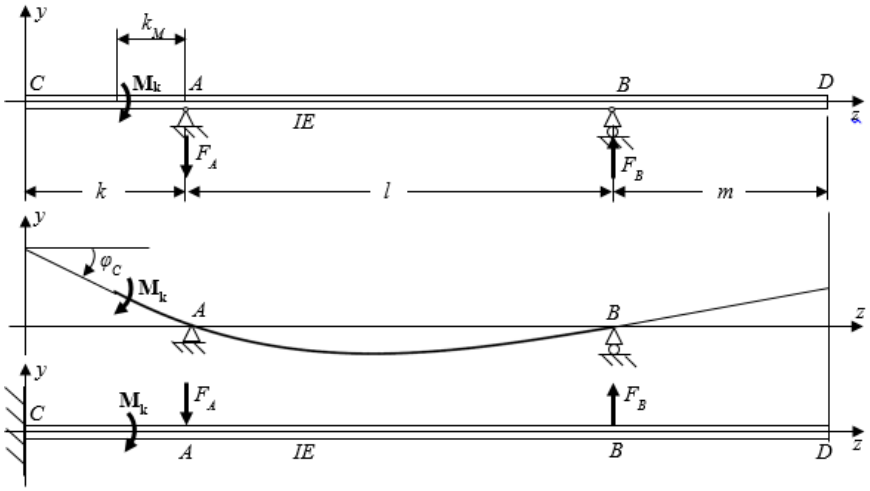


Figure 13

Simple supported beam loaded by couple of force on the left console at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section C

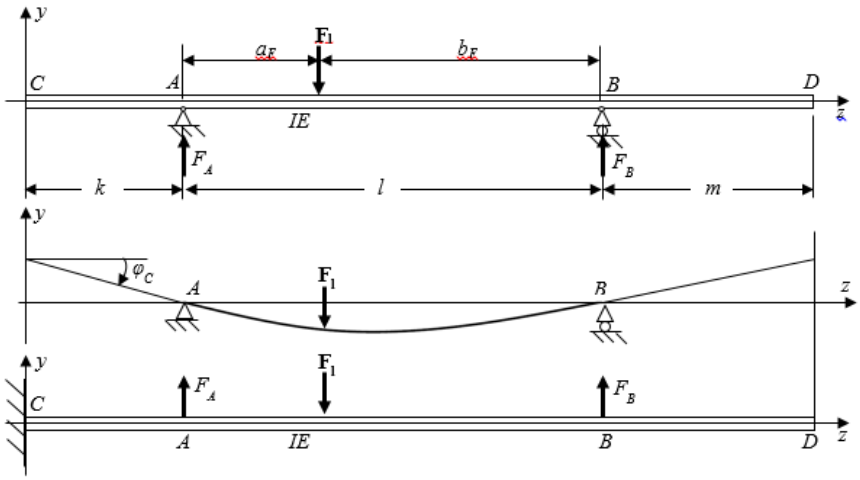


Figure 14

Simple supported beam loaded by concentrated force on effective span at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section

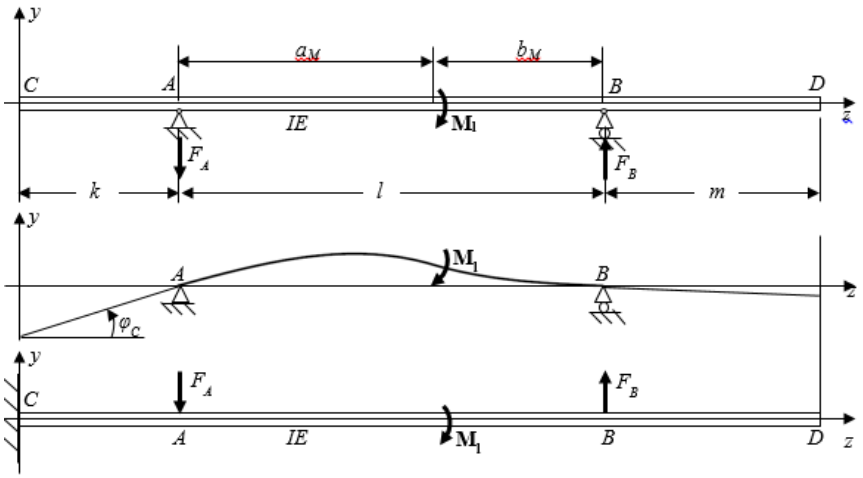


Figure 15

Simple supported beam loaded by couple of force on effective span at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section C

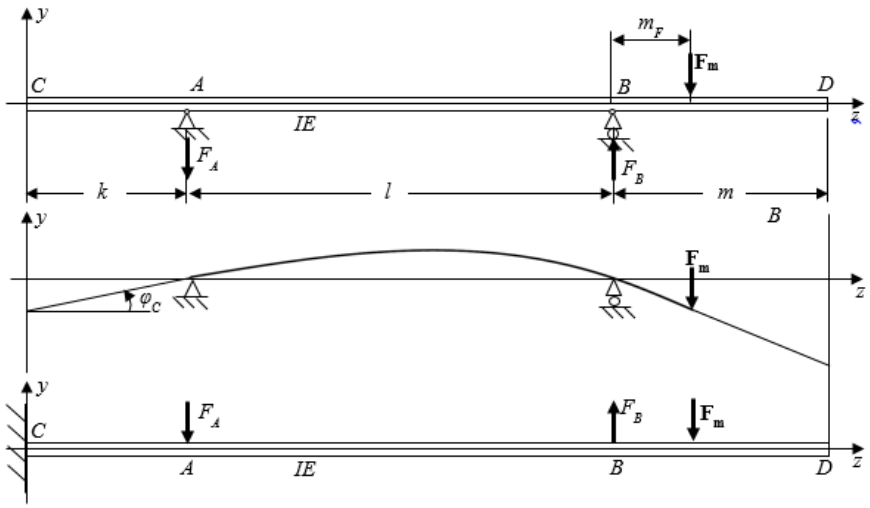


Figure 16

Simple supported beam loaded by concentrated force on the right console at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section C

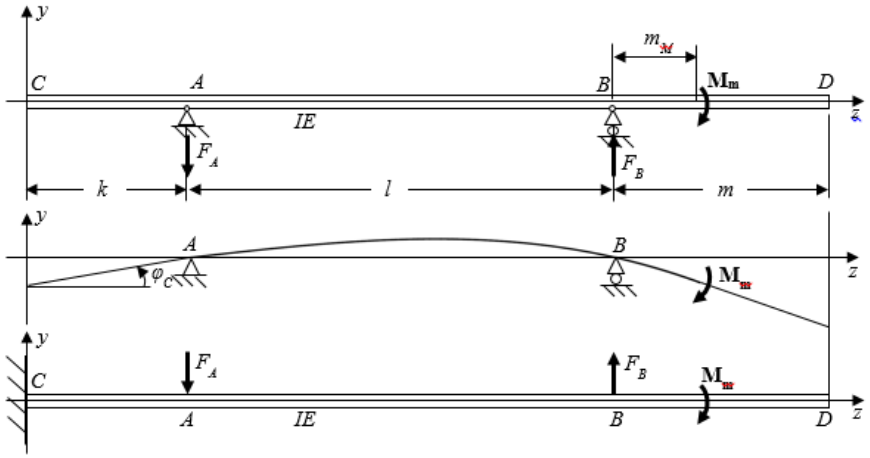


Figure 17

Simple supported beam loaded by couple of force on the right console at arbitrary place, shape of elastic curve (strong enlargement), moreover sketch of „modified" beam, i.e. clamped at cross-section C

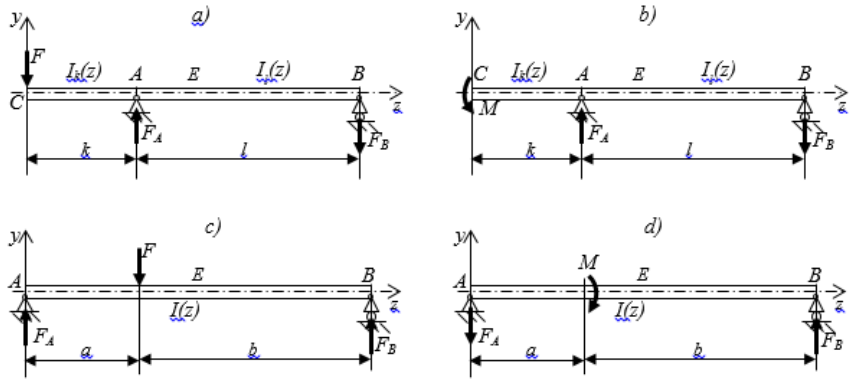


Figure 18 (a-d)

Simple supported beam loaded by concentrated force or couple of forces on the end of the cantilever or between its supports. The moment of inertia of the cross section about its neutral axis is continuous function of position coordinate z

References

[1] Ibhadode, O.; Dagwa, I.; Asibor, J.; Oho-Oghogho, E. Development of a computer aided deflection analysis (CABDA) program for simply supported loaded beams. *Int. J. Eng. Res. Africa* **2017**, *30*, 23-28, https://doi.org/10.4028/www.scientific.net/JERA.30.23

[2] Heng, Z.; Shu-Ying, Q.; Guo-Liang, W. Research on the method of simply supported beam modal parameters recognition by QY inclinometer. *J. Appl. Sci* **2014**, *14*, 1844-1850, https://doi.org/10.3923/jas.2014.1844.1850

[3] Yang, N.; Qin, S. Effect of quite-inclination angles on structural performance of Tibetan Timber beam-column joints. *J. Perform. Constr. Facil.* **2018** *32*, 12 p, https://doi.org/10.1061/(ASCE)CF.1943-5509.0001156

[4] Rojas, A. L.; Chavarria, S. L.; Elizonda, M. M.; Kalashnikov, V. V. A mathematical model of elastic curve for simply supported beams subjected to a concentrated load taking into account the shear deformations. *Int. J. Innov. Comput. Inf. Control., ICIC Exp. Lett., Part B: Appl.* **2016** *12*, 41-54

[5] Ramachandra, L. S.; Roy, D. The locally transversal linearization (LTL) method revisited: A simple error analysis *J. Sound Vib.* **2002,** *256*, 579-589, https://doi.org/10.1006/jsvi.2001.4222

[6] Kumar, R., Ramachandra, L. S., Roy, D. A multi-step linearization techniques for a class of bending value problems in non-linear mechanics *Comput. Mech.* **2006**, *39*, 73-81, https://doi.org/10.1007/s00466-005-0009-6

[7] Viswanath, A.; Roy, D. Multi-step transversal and tangential linearization methods applied to a class of nonlinear beam equations *Int. J. Solids. Struct.* **2007,** *44*, 4872-4891, https://doi.org/10.1016/j.ijsolstr.2006.12.008

[8] Merli, R.; Lázaro, S.; Monleón, S., Domingo, A. Comparison of two linearization schemes for the nonlinear bending problem of a beam pinned at both ends *Int. J. Solids. Struct.* **2007,** *47*, 865-874, https://doi.org/10.1016/j.ijsolstr.2009.12.001

[9] Thankane, K. S.; Stys, T. Finite difference method for beam equation with free ends using Mathematica *SAJPAM* **2009,** *4*, 61-78

[10] Bíró, I.; Cveticanin, L.; Szuchy. P. Numerical method to determine the elastic curve of simply supported beams of variable cross-section′ *Struct. Eng. Mech. Int. J.* **2018,** *68*, 713-720, https://doi.org/10.12989/sem.2018.68.6.713

[11] Abu-Alshaikh, I.; Alkhaldi, H.; Beithou, N. Large deflection of prismatic cantilever beam exposed to combination of end inclined force and tip moment *Mod. Appl. Sci.* **2018**, *12*, 98-111, https://doi.org/10.5539/mas.v12n1p98

[12] Batista, M. Large deflection of a beam subject to three-point bending *Int. J. Non-Linear Mech.* **2015,** *69*, 84-92, https://doi.org/10.1016 /j.ijnonlinmec.2014.11.024

[13] Abolfathi, A.; Brennan, M. J.; Waters, T. Large deflection of simply supported beam, *ISVR Technical Memorandum No. 988,* University of Southampton, **2010**

[14] Hatami, M.; Vahdani, S.; Ganji, D. D. Deflection prediction of a cantilever beam subjected to static co-planar loading by analytical methods *HBRC J.* **2014** *10*, 191-197, https://doi.org/10.1016/j.hbrcj.2013.11.003

[15] Batista, M. Analytical solution for large deflection of Reissner's beam on two supports subjected to central concentrated force *Int. J. Mech. Sci* **2016,** *107*, 13-20, https://doi.org/10.1016/j.ijmecsci.2016.01.002

[16] Machado, J. A. T.; Babaei, A.; Moghaddam, B. P. Highly Accurate Scheme for the Cauchy Problem of the Generalized Burgers-Huxley Equation. Acta Polytechnica Hungarica Vol. 13, No. 6, 2016, 183-195, DOI: 10.12700/APH.13.6.2016.6.10

[17] Horváth, L.; Rudas, I. J. Modeling and Problem Solving Techniques for Engineers. Elsevier Academic Press, 2004, London, ISBN 9780126022506

[18] Cveticanin, L.; Mester, Gy. Theory of Acoustic Metamaterials and Metamaterial Beams: An Overview. Acta Polytechnica Hungarica, Vol. 13, No. 7, 2016, 43-62, DOI: 10.12700/APH.13.7.2016.7.3

Control of Deadlocked Discrete-Event Systems Using Petri Nets

František Čapkovič

Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia, e-mail: Frantisek.Capkovic@savba.sk

Abstract: In Discrete-Event Systems (DES), deadlocks frequently occur. Flexible Manufacturing Systems (FMS) have the character of DES. Namely, FMS consist of many cooperating devices (like robots, machine tools, transport belts, etc.). Frequently, deadlocks occur because of insufficient resources. Petri Nets (PN) are often used to model FMS and to synthesize control for them. To deal with deadlocks, first of all, it is necessary to find and/or avoid them. There are several principal approaches for doing this - either by computing and analyzing the PN reachability tree (RT) or by finding PN model siphons. Then, in the former concept, the supervisor is synthesized by means of P-invariants of the PN model used, while in the latter concept the supervisor, based on siphons, is synthesized. In addition to these approaches, additional techniques can sometimes, be applied - e.g. a suitable utilization of added PN transitions.

Keywords: control; deadlocks; discrete-event systems; Petri nets; P-invariants; siphons

1 Introduction

In Discrete-Event Systems (DES) a next state depends only on the actual state and on the occurrence of discrete events. For modeling and control of DES Petri nets (PN) are frequently used. One of the typical representatives of DES is the family of Flexible Manufacturing Systems (FMS), newer Automated Manufacturing Systems (AMS). In such systems (robotized working cells, discrete production lines, and the like) many devices cooperate together - robots, machine tools, transport belts, automatically guided vehicles (AGV), etc. They are called to be *resources.* Inside FMS/AMS the resource allocation is very important. Hence, Resource Allocation Systems (RAS) are investigated.

For above mentioned reasons *deadlocks* often occur in RAS. Deadlocks are, of course, undesirable and unfavorable. They disrupt the normal course of the production process. Due to deadlocks, it remains stagnate. Thus, the primary intention of the production cannot be achieved. Deadlocks can arise, for example, when a machine M, completes a part and there is no part in a buffer to be fed to M,

it is the situation called *starvation*. In general, release of too few parts to RAS may starve some machines and lower their production rate. Therefore, it is necessary to pursue the maximally permissive control policy, for deadlock avoidance [1] by releasing as many jobs as possible into the system. On the other hand, when a machine M completes a part that cannot be unloaded because of the lack of buffer spaces, it is *blocked*. In general, *blocking* is caused by the excessive job releases and limited buffer spaces. Blocked machines are forced to be idle, thereby they loss their productivity. The more parts occur in the system, the more likely is occurrence of deadlocks and the machines are blocked. To operate RAS effectively, the system should be well scheduled and deadlocks should be completely avoided, in this way, the reduction of starvation and blocking is efficiently achieved.

To deal with the deadlocks, it is necessary to find them, and to find a suitable methodology for how to eliminate their impact and to successfully control the system. The deadlocks in FMS can be found by applying two main manners:

(i) Finding and analyzing the reachability tree (RT) of the PN model representing the causality of PN states.

(ii) Finding and analyzing structural properties - namely the set of siphons and traps of the PN model. Traps are some complements of siphons. Sometimes, the approach using an application of additional transition(s) into the PN model may be very suitable.

As to the control of deadlocked FMS, i.e. elimination of the deadlock impact, three concepts of synthesizing the supervisor are used here:

(i) The approach based on *P*-invariants of the PN model and simultaneous utilizing its RT

(ii) The approach-based on PN model siphons and traps

(iii) The auxiliary approach adding some supplementary transitions to the PN model

1.1 The State of the Art Review

Deadlocks are looking for and analyzed for tens of years in software engineering and other branches, and for a long time also in FMS/AMS.

Among pioneers of deadlock avoidance in DES and RAS belongs S. A. Reveliotis with his school. Their oldest publications were devoted especially to software engineering, but their newer ones - see [2-4] - are concerning RAS in FMS/AMS. Another school around P. J. Antsaklis - see [5] - is specialized on RT-based approaches to the deadlock avoidance. Both schools are American.

There are many other authors, even schools, interested in this area. The newer schools are the Chinese-American school around Meng Chu Zhou and Zhi Wu Li -

see [6-9] specialized on siphon-based approaches. This school publishes very intensively. Smaller schools are the French school around K. Barkaoui - see e.g. [10], and several others schools - in Spain [12], Italy [12-14], Germany [15].

But at present the peak school is the pure Chinese school with enormous number of authors from various universities - see e.g. [16-18] - with top results. Some authors from the above introduced schools publish also together cross by cross the schools. It is impossible to make a complete overview of all the works of these schools on a limited number of pages per paper.

Deadlocks may occur also in DES and RAS with non-determinism analysed in the paper [19]. Such deadlocks have to be avoided too.

Simply, this area of research in FMS/AMS lives through a boom. Therefore, it is useful to choose the more important approaches and compare them. The best form is to do this by applying them on the same real plant and compare and evaluate their results. Such a process has not be published until now.

1.2 The Main Aim of the Paper

The main aim of this paper is to point out:

(i) The three principal kinds of approaches extracted from the huge amount of literature

(ii) How to avoid deadlocks in real RAS

(iii) How to synthesize the control of RAS by means of PN models

(iv) How to apply particular approaches on the same real discrete plants in order to compare them

(v) How to perform the comparison on the basis of achieved results and how they are evaluated. (In the literature, the author of this paper did not find such comparison and evaluation of different approaches).

Of course, finding the computational complexity of algorithms for computing RT (at the state analysis) and minimal siphons (at the structural analysis), respectively, is also an associated, but not less important, aim.

For FMS/AMS practice such a comparison and evaluation may be very useful. Namely, on one hand it is important to avoid deadlocks, but on the other hand it is also necessary to detect whether the proposed supervisor avoiding deadlocks ensures satisfying functionality of RAS or not. If not, a structural reconstruction of the original system and/or changing the number of resources and repeating the procedure of the supervisor synthesis is needful.

1.3 Paper Organization

Here, in this Section 1, the state of the art review and the main aim of this paper are introduced. In the Section 2, the PN themselves as well as PN *P*-invariants and PN siphons and traps were defined and two approaches how to control RAS were sketched. In the next Section 3 the auxiliary simple approach to control of deadlocked RAS will be introduced and illustrated on an example, namely the approach based on additional (supplementary) transitions. Next, in the Section 4, the second approach to solving that problem will be introduced and illustrated on examples, namely the approach based on *P*-invariants. In the Section 5 the approach based on siphons and traps will be presented and illustrated on examples. In the Section 6 both approaches will be compared (as to their advantageous and disadvantageous) and evaluated. In the Conclusions, the final view on the dealing with deadlocks in this paper and the plans for future research will be introduced.

2 Preliminaries

PN are perspective tool [12] for modeling and DES. Essentials of PN were presented in many older papers - see e.g. [10]. The state equation of place/transition PN (P/T PN) - see [19] - is the following

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{B}.\mathbf{u}_k \ , k = 0, 1, \ldots \tag{1}$$

$$\mathbf{F}.\mathbf{u}_k \leq \mathbf{x}_k \tag{2}$$

where

$\mathbf{x}_k = (x_{p_1}, x_{p_2}, \ldots, x_{p_n})^T$ is the state vector with integer entries $x_{p_i} \in \{0, 1, \ldots, \infty\}$ being the states of particular places p_i, i = 1, …, n, in the step k, namely x_{p_i} represents the actual number of tokens in the place p_i. The vector $\mathbf{x}_0$ is the initial state vector.

$\mathbf{u}_k = (u_{t_1}, u_{t_2}, \ldots, u_{t_m})^T$ is the control vector with entries $u_{t_j} \in \{0, 1\}$ being the states of particular transitions t_j, $j = 1, \ldots, m$, in the step k. They can be disabled or enabled. The disabled t_j cannot be fired, i.e. $u_{t_j} = 0$, while enabled t_j may be (but needs not be) fired i.e. $u_{t_j} = 1$. In P/T PN enabled transitions represent the occurrence of discrete events.

$\mathbf{B} = \mathbf{G}^T - \mathbf{F}$ is the structural matrix of integers with **G** being the incidence matrix of directed arcs from transitions to places while **F** being the incidence matrix of directed arcs from places to transitions

Let $P = \{p_1, \ldots, p_n\}$ and $T = \{t_1, \ldots, t_m\}$ are, respectively, the set of PN places and the set of PN transitions. Thus, $\mathbf{F} = \{f_{ij}\}$, $i = 1, \ldots, n$; $j = 1, \ldots, m$, $f_{ij} \in Z$, where Z is the set of integers, and it represents the existence and multiplicity of arcs directed

from p_i to t_j; $\mathbf{G} = \{g_{ji}\}, j = 1, \ldots, m; i = 1, \ldots, n, g_{ji} \in Z$, and it represents the existence and multiplicity of arcs directed from t_j to p_i.

$\mathbf{x}_0$ is the initial state vector.

Starting from $\mathbf{x}_0$ and firing an enabled transition the next state $\mathbf{x}_1$ can be reached. The reachability tree (RT) expresses all possible branches of the development of the system (1), (2). A firing sequence of transitions t_a, t_b, …,t_c represents a branch $\mathbf{x}_0 \xrightarrow{u_{t_a}} \mathbf{x}_1 \xrightarrow{u_{t_b}} \ldots \mathbf{x}_{k-1} \xrightarrow{u_{t_c}} \mathbf{x}_k$ of RT. All reachable states create the state space, i.e. the set $\mathcal{R} = \{ \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k \}$. PN transitions symbolize edges of RT. By means of the thorough analysis of RT (either in graphical form or in the form of the adjacency matrix) all deadlocks can be found. Then the supervisor based on *P*-invariants can be synthesized. The *P-invariant* is the (n×1)-dimensional vector $\mathbf{y} \neq \mathbf{0}$ for which $\mathbf{y}^T.\mathbf{B} = \mathbf{0}$.

A nonempty subset $S \subset P$ in P/T PN is called a *siphon* if every transition having an output place in S has an input place in S. A nonempty subset $Q \subset P$ in P/T PN is called a *trap* if every transition having an input place in Q has an output place in Q. Siphons create a set of places which, if become empty of tokens, will always *remain empty* for all reachable markings of the net. When all places in a siphon have no token, all transitions connecting with the siphon can no longer be firable. Traps create a set of places which, if become marked, will always *remain marked* for all reachable markings of the net. The union of two siphons (traps) is again a siphon (trap).

If every non-empty siphon of PN includes a sufficiently marked trap then - see e.g. [6], no dead marking is reachable. It is very important piece of knowledge. Thorough analysis of siphons and traps is a path to the proposal of the supervisor. Then, the supervisor will be synthesized by means of utilizing properties of siphons and traps.

It is not necessary to work with all siphons (there are many). It is sufficient to work with *elementary siphons*, i.e. linearly independent siphons. Even, it is sufficient to work with *minimal siphons* and *minimal traps* - see e.g. [11] [20].

In next, both approaches to control of deadlocked FMS, especially of the special kind of FMS called RAS, will be presented. Namely, above mentioned devices - machine tools, robots, buffers, transport belts, AGV, and so on, can be understood to be various resources. The resources are usually shared by two or more subsystems of RAS. Because of a limited number of resources different kinds of problems, especially deadlocks, arise during the system operation.

3 Approach Based on Additional Transitions

This approach is very useful especially in the case of so called *diamonds* in RT of the deadlocked PN model. The diamond ∂ from the start state $\mathbf{x}_a$ in RT to the end state $\mathbf{x}_b$, $\mathbf{x}_a \neq \mathbf{x}_b$, is a pair of paths $\partial = <\mathbf{x}_a\ \sigma_1\ \mathbf{x}_b, \mathbf{x}_a\ \sigma_2\ \mathbf{x}_b>$, where paths $\sigma_1 \cap \sigma_2 = \emptyset$, $\sigma_1 \cup \sigma_2 \neq \emptyset$, with $\emptyset$ being the empty set, and $\mathbf{x}_a$, $\mathbf{x}_b$ do not belong in $\sigma_1 \cup \sigma_2$. When $\mathbf{x}_b$ is the deadlock, the following approach is possible in order to deal with it. After adding a transition into the PN model, the structural matrix of the supervisor:

$$\mathbf{B}_c = \mathbf{x}_c - \mathbf{x}_b \tag{3}$$

where usually $\mathbf{x}_c = \mathbf{x}_0$ (being the initial state of the PN model (1)-(2)).

3.1 Example 1

Consider the RAS in Figure 1 consisting of three loading buffers I1–I3 and three unloading buffers O1–O3. They, respectively, load and unload the FMS corresponding to three raw product types, Pr1–Pr3, to be processed by machine M. They are moved by robot R. The production cycles are the following: a raw product Pr1 is taken from I1 by R and put in M. After being processed by M, it is taken by R and put to output O1. A raw product Pr2 is taken by R from I2, processed by M and then moved by R from M to O2. A raw product Pr3 is taken by R from I3, processed by M and moved by R from M to O3. The PN model is given in Fig. 2. The initial state $\mathbf{x}_0 = (8\ 0\ 0\ 8\ 0\ 0\ 1\ 0\ 4\ 4)^T$. Because figures of RT from $\mathbf{x}_0$ produced by a graphic tool has a poor quality at a greater dimensionality of RT, the deadlocks will be computed by means of the zero rows of the RT adjacency matrix. There are 32 nodes (states) in the RT.

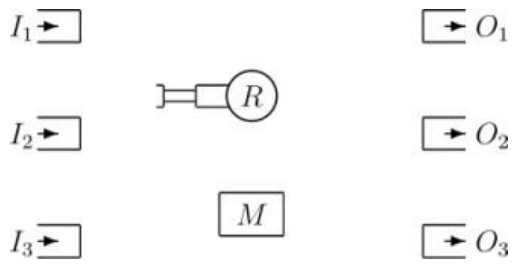


Figure 1
The scheme of RAS

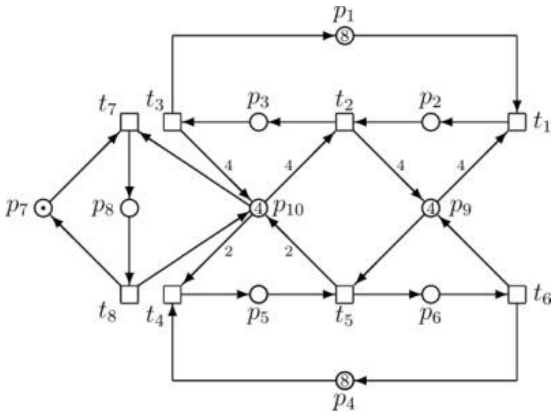


Figure 2
The PN model of the deadlocked RAS

The structural matrix of the PN model and its initial state are the following:

$$\mathbf{B} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ -4 & 4 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & -4 & 4 & -2 & 2 & 0 & -1 & 1 \end{pmatrix}$$

$\mathbf{x}_0 = (8\ 0\ 0\ 8\ 0\ 0\ 1\ 0\ 4\ 4)^T$

Small numbers in the neighborhood of some directed arcs mean their multiplicity.

This plant has only one deadlock – the state $\mathbf{x}_{11} = (7\ 1\ 0\ 6\ 2\ 0\ 1\ 0\ 0\ 0)^T$ (i.e. numbers of tokens in corresponding places are: $p_1 = 7$, $p_2 = 1$, $p_4 = 6$, $p_5 = 2$, $p_7 = 1$) is the deadlock. Thus, $\mathbf{B}_c = \mathbf{x}_0 - \mathbf{x}_{11} = (1\ -1\ 0\ 2\ -2\ 0\ 0\ 0\ 4\ 4)^T$, where the vector $\mathbf{B}_c^{(-)} = (0\ 1\ 0\ 0\ 2\ 0\ 0\ 0\ 0\ 0)^T$ represents the multiplicity of directed arcs from corresponding PN places to the added transition and the vector $\mathbf{B}_c^{(+)} = (1\ 0\ 0\ 2\ 0\ 0\ 0\ 0\ 4\ 4)^T$ represents those from the added transition to corresponding PN places. The PN model of the modified structure is given in Figure 3. Alike as in previous case, the RT is not introduced in the graphical form. From the RT adjacency matrix, it follows that PN has 32 nodes and no deadlock exists in the supervised system displayed in Fig. 3. Adding the transition t_9 into the PN model of RAS given in Figure 2 and applying its interconnections with the original model through the relation (3) the deadlock $\mathbf{x}_{11}$ was eliminated.

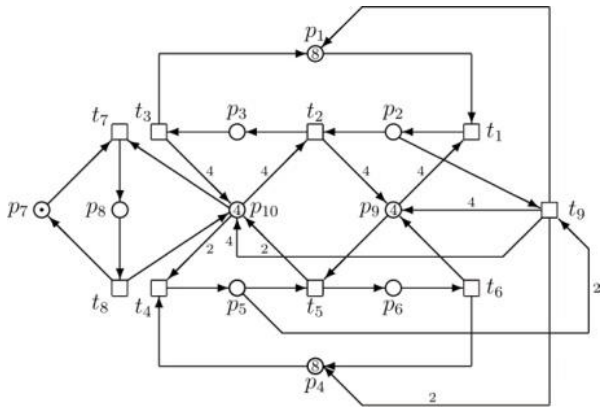


Figure 3
The PN model of RAS supervised by means of the transition t_9 without any deadlock

4 Approach Based on P-invariants

As it was mentioned above a vector $\mathbf{y} \neq \mathbf{0}$ fulfilling the relation $\mathbf{y}^T.\mathbf{B} = \mathbf{0}$ is named as the invariant. For more invariants, e.g. s, the matrix $\mathbf{Y}$ consisting of s invariants being its columns, has to fulfill:

$$\mathbf{Y}^T.\mathbf{B} = \mathbf{0} \tag{4}$$

Putting a restricted condition:

$$\mathbf{L}.\mathbf{x} \leq \mathbf{b}, \tag{5}$$

where $\mathbf{L}$ is a $(s \times n)$ matrix of positive integers, expressing by its rows suitable linear combinations of state vectors entries (i.e. numbers of tokens inside corresponding PN places), and $\mathbf{b}$ is a $(s \times 1)$ column vector of limits for each row of $\mathbf{L}$ (i.e. a maximal number of tokens in places in the corresponding row together). To remove inequality in (4), we can put the following:

$$\mathbf{L}.\mathbf{x} + \mathbf{x}_s = \mathbf{L}.\mathbf{x} + \mathbf{I}_s.\mathbf{x}_s = (\mathbf{L}\ \mathbf{I}_s).(\mathbf{x}^T\ \mathbf{x}_s^T)^T = \mathbf{b} \tag{6}$$

where $\mathbf{I}_s$ is the $(s \times s)$ identity matrix. To synthesize the supervisor with the structure $\mathbf{B}_s$ (unknown till now), we force $(\mathbf{L}\ \mathbf{I}_s)$ into (4) instead of $\mathbf{Y}^T$ as well as $(\mathbf{B}^T\ \mathbf{B}_s^T)^T$ instead of $\mathbf{B}$. In such a way we finally obtain the supervisor structure:

$$\mathbf{B}_s = -\mathbf{L}.\mathbf{B} \tag{7}$$

$$\mathbf{B}_s = \mathbf{G}_s^T - \mathbf{F}_s \tag{8}$$

$$\mathbf{F}_s = -\mathbf{B}_s^{(-)};\ \mathbf{G}_s^T = \mathbf{B}_s^{(+)}$$

$$\mathbf{x}_{0s} = \mathbf{b} - \mathbf{L}.\mathbf{x}_0 \tag{9}$$

Besides (5) the *general linear constraints* can be imposed to be satisfied by the supervised system:

$$\mathbf{L}_p.\mathbf{x} + \mathbf{L}_t.\mathbf{u} + \mathbf{L}_v.\mathbf{v} \leq \mathbf{b} \tag{10}$$

where,

$\mathbf{b}$ is s - dimensional nonnegative integer vector expressing some limits

$\mathbf{L}_p$, $\mathbf{L}_t$, $\mathbf{L}_v$ are, respectively, $(s \times n)-$, $(s \times m)-$, $(s \times m)-$ dimensional matrices of integers. They concern, respectively, PN places, PN transitions and the Parikh's vector $\mathbf{v}$. The sense of the Parikh's vector $\mathbf{v}$ is clear from the following relation expressing the evaluation of PN model (1), (2), i.e.

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{B}.(\mathbf{u}_0 + \mathbf{u}_1 + \ \ldots\ + \mathbf{u}_{k-1}) = \mathbf{x}_0 + \mathbf{B}.\mathbf{v} \tag{11}$$

As to (11), it was proved in [5] that when $\mathbf{L}_p.\mathbf{x} - \mathbf{b} \leq \mathbf{0}$ the supervisor with the following structure and initial state

$$\mathbf{F}_s = \max(\mathbf{0}, \mathbf{L}_p.\mathbf{B} + \mathbf{L}_v, \mathbf{L}_t) \tag{12}$$

$$\mathbf{G}^{\mathrm{T}}{}_s = \max(\mathbf{0}, \mathbf{L}_t - \max(\mathbf{0}, \mathbf{L}_p.\mathbf{B} + \mathbf{L}_v)) - \min(\mathbf{0}, \mathbf{L}_p.\mathbf{B} + \mathbf{L}_v) \tag{13}$$

$$\mathbf{x}_{0s} = \mathbf{b} - \mathbf{L}_p.\mathbf{x}_0 - \mathbf{L}_v.\mathbf{v}_0 \tag{14}$$

guarantees that constraints are verified for the states resulting from the initial state $\mathbf{x}_0$. Here, the max(.) is the maximum operator of operands. For matrices it is applied element by element, i.e. $\mathbf{Z} = \max(\mathbf{X}, \mathbf{Y})$ means that $z_{ij} = \max(x_{ij}, y_{ij})$.

$\mathbf{v}_0$ is the $(m \times 1)$ vector containing nonzero entries (namely equal to 1) solely in positions of transitions being firable in $\mathbf{x}_0$.

Now, consider the RAS in Figure 1 with PN model given in Figure 2. Let us deal with the deadlock state $\mathbf{x}_{11}$ using the *P*-invariants based approach for both versions the simpler (5) and the generalized (10).

4.1 Simpler Version of the Approach

Analyzing RT by means of the adjacency matrix we can reveal that the deadlocked state $\mathbf{x}_{11}$ is reached by two ways:

$$\mathbf{x}_0 \xrightarrow{t1} \mathbf{x}_1 \xrightarrow{t4} \mathbf{x}_5 \xrightarrow{t4} \mathbf{x}_{11} \tag{15}$$

$$\mathbf{x}_0 \xrightarrow{t4} \mathbf{x}_2 \xrightarrow{t4} \mathbf{x}_7 \xrightarrow{t1} \mathbf{x}_{11} \tag{16}$$

The corresponding critical RT nodes are the following:

$\mathbf{x}_5 = (7\ 1\ 0\ 7\ 1\ 0\ 1\ 0\ 0\ 2)^T$

$\mathbf{x}_7 = (8\ 0\ 0\ 6\ 2\ 0\ 1\ 0\ 4\ 0)^T$

$\mathbf{x}_{11} = (7\ 1\ 0\ 6\ 2\ 0\ 1\ 0\ 0\ 0)^T$

$\mathbf{x}_5 - \mathbf{x}_{11} = (0\ 0\ 0\ 1\ -1\ 0\ 0\ 0\ 0\ 2)^T$

$\mathbf{x}_7 - \mathbf{x}_{11} = (1\ -1\ 0\ 0\ 0\ 0\ 0\ 0\ 4\ 0)^T$

Using the approach (6)-(8) based on *P*-invariants, we have to put some restrictions on combinations of some critical places. Such places are p_2, p_5 and moreover p_8, p_{10}. Namely, the following restrictions have to be imposed:

$$p_8+p_{10} \leq 5 \tag{17}$$

$$p_2+p_5 \leq 1 \tag{18}$$

Consequently,

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{19}$$

$$\mathbf{b} = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \tag{20}$$

Applying these matrices into (6)-(8) we obtain the supervisor structure and initial state as follows:

$$\mathbf{B}_s = \begin{pmatrix} 0 & 4 & -4 & 2 & -2 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{F}_s = -\mathbf{B}_s^{(-)} = \begin{pmatrix} 0 & 0 & 4 & 0 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{G}_s^T = \mathbf{B}_s^{(+)} = \begin{pmatrix} 0 & 4 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{x}_{0s} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The PN model of RAS controlled by such supervisor is displayed in Figure 4. The supervisor ensures that no deadlock occurs here. RT has 24 nodes in this case.

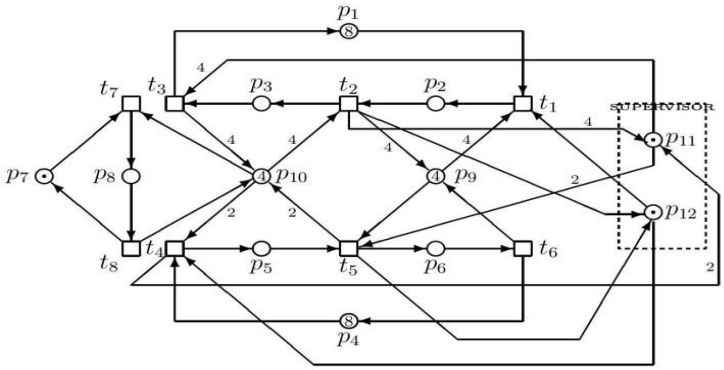


Figure 4

The PN model of RAS controlled by the *P*-invariants based supervisor removing the deadlock

After deeper analysis, we can found that the condition (17) may be omitted and it is sufficient in this RAS to use solely the condition (18). Thus, we obtain:

$\mathbf{L}$= (0 1 0 0 1 0 0 0 0 0)

$\mathbf{b}$ = (1)

consequently,

$\mathbf{F}_s$= (1 0 0 1 0 0 0 0)

$\mathbf{G}^T_s$= (0 1 0 0 1 0 0 0)

$\mathbf{x}_{0s}$ = (1)

The PN model of the controlled RAS is almost-certain as that on Figure 4, only place p_{11} together with its interconnections are missing. However, the RT is the same. It means that no deadlocks are indicated in it.

4.2 The Generalized Version of the Approach

Here we use the approach (10)-(14) to illustrate its potency. From the PN model of the original uncontrolled system in Figure 2 it can be seen that main problem consists in the places p_2 and p_5 as well as in Parikh's vectors v_1 and v_4. Let us put $p_2 + p_5 \leq 1$ and because of the Parikh's vectors put $t_4 > t_1$ and $t_1 > t_4$. Hence:

$$\mathbf{L}_p = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathbf{L}_v = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}; \mathbf{L}_t = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{v}_0 = (1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0)^T$ is the initial vector expressing enabled transitions at the initial state $\mathbf{x}_0$.

These inputs into (12) - (14) result the following synthesized supervisor

$$\mathbf{F}_s = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{G}_{ts} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{x}_{0s} = (1 \ 1)^T$

The PN model of the controlled RAS is displayed in Figure 5 and its RT is given in Figure 6. For such a small RT like this, as opposed to greater ones, the graphical output of RT has a better quality, is readable and sufficiently highlighted. No deadlock occurs there.

As to evolution of the PN model behavior, restrictions on the supervised system are more rigorous. Thus, the RT is not so much branching out but in spite of this it guarantees a deadlock-free behavior of the PN model. It has only 13 nodes

(including the initial node) - states of the system. In comparison with the previous one given in Figure 4, having 24 nodes, it can be seen that this structure of controlled RAS yields only about half number of states, what may be insufficient from practical point of view (e.g. the functionality of a real plant, its utility, etc.). The user in practice has to consider if this structure is adequate for his requirements, or he will use the previous structure displayed in Figure 4. Namely, the too severe supervisor can hamper the required behavior of the system. In such a case the user has to change the structure of RAS and/or the number of resources and to repeat the whole process of the supervisor synthesis.

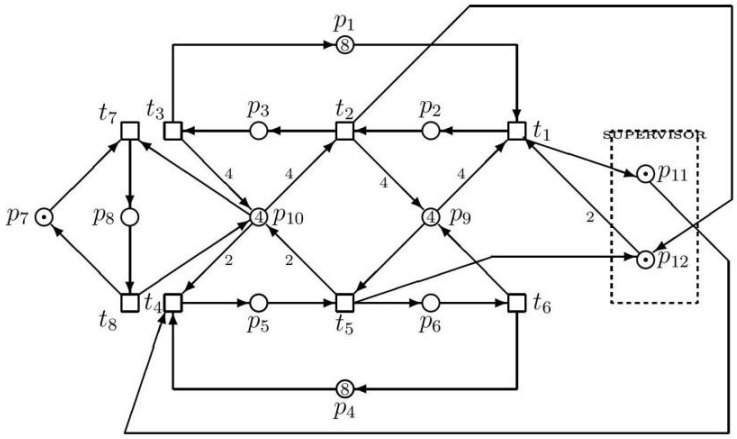


Figure 5
The PN model of the controlled RAS

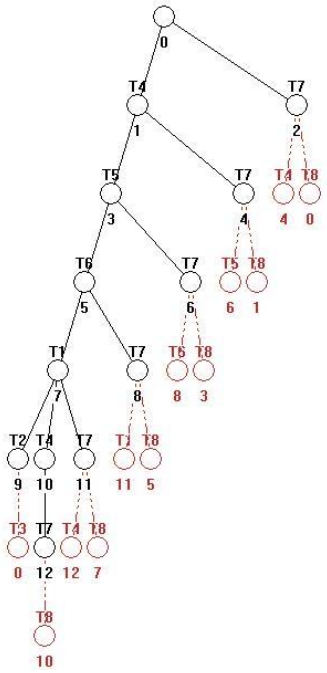


Figure 6
The RT of the PN model of the controlled RAS

4.3 Example 2

Consider RAS schematically sketched in Figure 7.

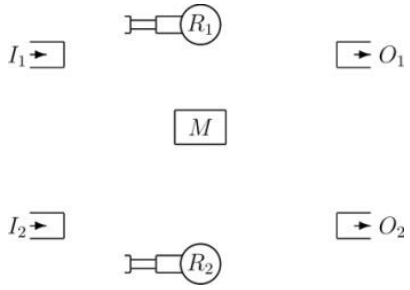


Figure 7
The scheme of RAS

There are two uploading buffers I_1 and I_2 and two unloading buffers O_1 and O_2 in order to upload and download RAS by two raw product types $Pr1$ and $Pr2$. They are processed by machine M and moved by robot R_1 (resp. R_2) and R_2 (resp. R_1). There are two production cycles: (i) a raw product $Pr1$ is taken from I_1 by R_2 and put in M. After being processed by M, the product is unloaded by R_1 and put to O_1.

(ii) a raw product $Pr2$ is taken from I_2 by R_1 and put in M. After being processed by M it is moved from M to O_2 by R_2.

The PN model of this RAS is given in Figure 8. RT of the PN model is too large for displaying here, because it has 216 nodes (including $\mathbf{x}_0$). It (more precisely its adjacency matrix) points out on 5 deadlocks - $\mathbf{x}_{54}$, $\mathbf{x}_{58}$, $\mathbf{x}_{62}$, $\mathbf{x}_{86}$, $\mathbf{x}_{121}$.

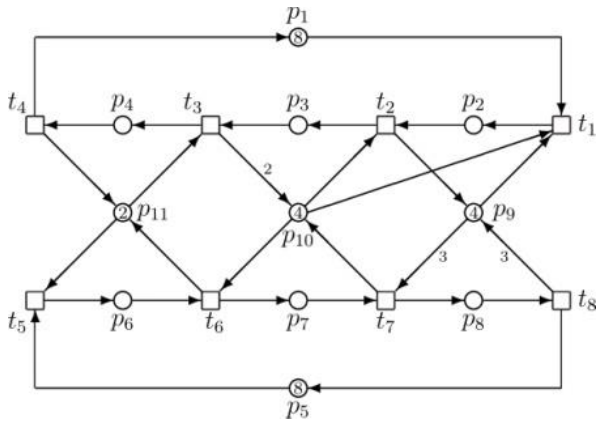


Figure 8
The PN model of RAS

4.3.1 P-invariant Based Approach

Let us apply the approach based on P-invariants to resolve the problem of deadlock avoidance.

$\mathbf{x}_0 = (8\ 0\ 0\ 0\ 8\ 0\ 0\ 0\ 4\ 4\ 2)^T$; $\mathbf{x}_{54} = (4\ 4\ 0\ 0\ 6\ 2\ 0\ 0\ 0\ 0\ 0)^T$

$\mathbf{x}_{58} = (5\ 2\ 1\ 0\ 6\ 2\ 0\ 0\ 2\ 0\ 0)^T$; $\mathbf{x}_{62} = (6\ 0\ 2\ 0\ 6\ 2\ 0\ 0\ 4\ 0\ 0)^T$

$\mathbf{x}_{86} = (5\ 3\ 0\ 0\ 5\ 2\ 1\ 0\ 1\ 0\ 0)^T$; $\mathbf{x}_{121} = (6\ 2\ 0\ 0\ 4\ 2\ 2\ 0\ 2\ 0\ 0)^T$

The structural matrix and $\mathbf{x}_0$ of the original uncontrolled system are the following:

$$\mathbf{B} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ -4 & 4 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & -4 & 4 & -2 & 2 & 0 & -1 & 1 \end{pmatrix}$$

$\mathbf{x}_0 = (8\ 0\ 0\ 0\ 8\ 0\ 0\ 0\ 4\ 4\ 2)^T$

After analysis between deadlocks and relative states (nodes RT) we can put **L** and **b** as follows:

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 12 \end{pmatrix}$$

Hence, we obtain the structure and the initial state of the supervisor:

$$\mathbf{F}_s = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}; \quad \mathbf{G}_s{}^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{x}_{s0} = (2\ \ 1\ \ 2\ \ 1\ \ 4)^{\mathrm{T}}$

The PN model of the supervised system is displayed in Figure 9. No deadlocks occur there in the supervised system.

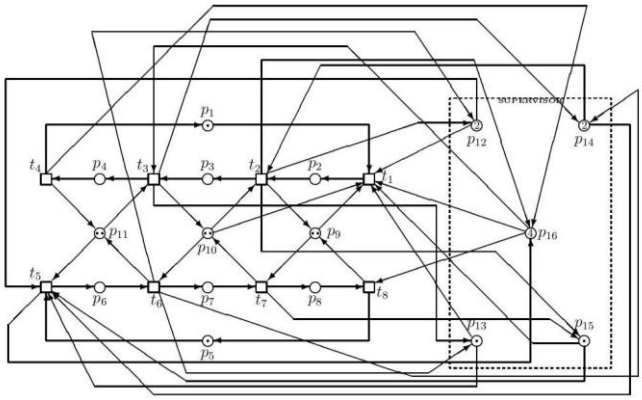


Figure 9
The PN model of the controlled RAS

5 Approach Based on Siphons and Traps

At this approach minimal siphons and minimal traps are computed. It may be realized e.g. using the tool [20] in Matlab. However, in general, at more complicated structure of the PN model and a big number of PN places, such approach may be also fairly time-consuming. This is valid for this tool too.

When we have minimal siphons in the matrix form $\mathbf{S}_M$ (with particular siphons being its rows), we can obtain the supervisor structure as follows:

$$\mathbf{B}s = \mathbf{S}_M.\mathbf{B}; \quad \mathbf{F}_s = -\mathbf{B}_s^{(-)}; \quad \mathbf{G}_s^T = \mathbf{B}_s^{(+)} \tag{21}$$

5.1 Application on Example 1

First of all, let us apply this approach on Example 1 in the Subsection 3.1. Let us resolve the problem with the deadlock in it by this way. The minimal siphons are the rows of the following matrix:

$$\mathbf{S}_M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Consequently, using the structural matrix $\mathbf{B}$ of the original system we obtain the structural matrix of the supervisor:

$$\mathbf{B}_s = \mathbf{S}_M.\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 3 & -1 & 1 & 0 & 0 & 0 \\ -4 & 1 & 3 & -2 & 2 & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{x}_{0s} = (0 \quad 0 \quad 3 \quad 0 \quad 3 \quad 4)^T$

The PN model of the controlled RAS is given in Figure 14.

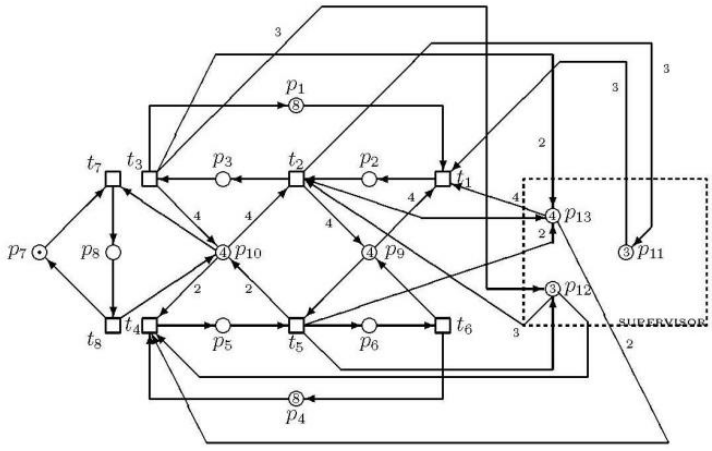


Figure 10
The PN model of the controlled RAS

5.2 Application on Example 2

Now, apply such an approach on the Example 2 analyzed in the Subsection 4.3. There are 199 siphons in the uncontrolled PN model and 8 minimal siphons (rows of the matrix $\mathbf{S}_M$) and 8 minimal traps (rows of $\mathbf{T}_M$), namely:

$$\mathbf{S}_M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{T}_M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

As we can see at comparing $\mathbf{S}_M$ an $\mathbf{T}_M$, minimal siphons $S_1 = Tr_1$, $S_3 = Tr_2$, $S_4 = Tr_3$, $S_7 = Tr_6$. Because these traps are marked and they cannot lose tokens, the corresponding siphons cannot stay deadlocks. It means, that at synthesizing of the supervisor it is sufficient to use only siphons as follows:

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Then, the structural matrix of the supervisor S is as follows:

$$\mathbf{B}_s = \mathbf{G}^{\mathrm{T}}{}_s - \mathbf{F}_s = \mathbf{S}.\mathbf{B} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 2 & 0 & -1 & 1 & 0 & 0 \\ -2 & 1 & 1 & 0 & 0 & -1 & -1 & 2 \\ -2 & 0 & 2 & 0 & -1 & 0 & -1 & 2 \end{pmatrix}$$

where $\mathbf{G}^{T}{}_s = \mathbf{B}_s{}^{(+)}$), while $\mathbf{F}_s = |\mathbf{B}_s{}^{(-)}|$).

For the initial state $\mathbf{x}_0$ of the uncontrolled PN model the initial state of the supervisor is $\mathbf{x}_{0s} = (4\ 6\ 8\ 10)^{\mathrm{T}}$. The PN model of the supervised system is displayed in Fig. 11. To verify if the supervised system is deadlock-free, we can compute RT starting from $\mathbf{x}_0$. It has 95 nodes. No deadlocks were found in RT.

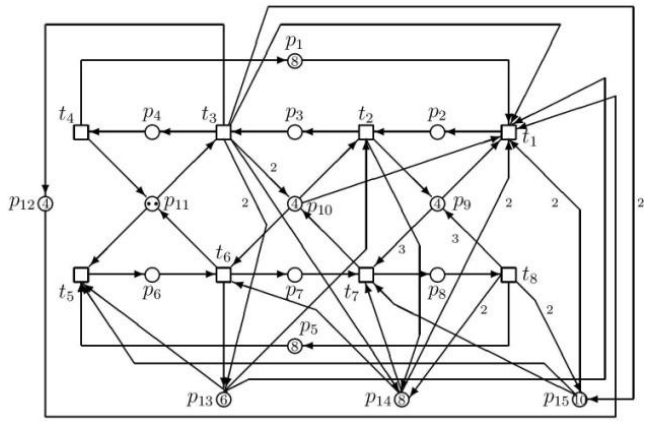


Figure 11
The PN model of the controlled RAS. The controller is created by the four places p_{12} - p_{15}

6 Comparison of the Presented Approaches

The approach in Section 3, using additional transitions, is only an auxiliary approach, but it has its importance, in cases with several diamonds in RT. Both principal approaches, presented in Section 4 and Section 5, respectively, are general and have their advantages and disadvantages. Moreover, the *P*-invariant based approach and the siphons and traps based one are much more powerful than that in Section 3. Therefore, only these two approaches will be compared here. The *P*-invariant based approach is more detailed because it works with PN places at the formulation of the matrix **L** and the vector **b**. Suitable combinations of places and assigning them maximal common numbers of tokens is very useful. On the other hand, at large or complicated PN models the computing of RT may be time-consuming and analyzing RT may be complicated. The siphons and traps based approach, does not need to compute RT, but only the siphons and traps have to be computed. However, the computation of them also depends on the size and especially on the complicacy of the PN model structure. This process can also be time-consuming, even more than that in previous case. Both of the compared approaches are very useful at synthesizing supervisor for deadlocked DES and especially RAS. Thus, they are able to avoid deadlocks and simultaneously, successfully control RAS.

Nevertheless, it is necessary to take into account the computational complexity of algorithms in both approaches, especially in the case of greater number of places and complicated structure of the PN model. The complexity at computation RT is between the upper bound $2^{O(\mathrm{n}\log n)}$ and the lower bound $2^{O(\sqrt{n})}$, where n is a number of PN places. At computing minimal siphons, the complexity is $O(2n + k(k-1)/2)$,

where n and k denote, respectively, the number of PN places and the number of all PN siphons. However, all PN siphons can be found with the computational complexity $O(2^n)$. Of course, the computational time depends on the hardware ability of a computer in question, however, the computational complexity of algorithms is unchanged.

Conclusions

With regard to the aims declared in Subsection 1.2, three approaches for control deadlocked DES were introduced in this paper. One of them, using insertion of additional transitions, is in effect, auxiliary but sometimes useful. The further two approaches, are very useful for avoiding deadlocks in DES and simultaneously for controlling them. For two examples, all approaches were applied and illustrated. It can be said that both of the essential approaches are very appropriate. Comparisons for them were also introduced. To declare unambiguously, which is better, further research is needed, especially testing larger and more complicated structures of DES and thus, more knowledge can be obtained. From the introduced examples, is clear, that the siphon and traps based approach, is more practical, since it employs one-stage and does not need any further computations. On the other hand, the RT based approach, uses two-stages. It needs deep analysis of RT and then, to set corresponding conditions, for the matrix **L** creation. A large RT cannot be seen in graphical form but only in the form of its adjacency matrix, which is also very large. Therefore, the creation of the matrix **L** may be time consuming and even sometimes, impossible. Conversely, the computational complexity of the siphon based approaches, is better at calculating RT. In a large RAS, it can lead to the long computational times. Hence, in general, new approaches with smaller computational complexity have to be explored.

As to our plans for further research, the utilization of the mixed integer programming (MIP) will be tested in the siphon based approach, for the control of RAS, since the interesting area seems to be discovering robust algorithms for the RAS control.

Acknowledgement

This work was partially supported by the Slovak Grant Agency for Science VEGA under Grant No. 2/0020/21.

References

[1] J. C. Luo, Z. Q. Liu, M. C. Zhou: A Petri Net Based Deadlock Avoidance Policy for Flexible Manufacturing Systems with Assembly Operations and Multiple Resource Acquisition. IEEE Transactions on Industrial Informatics, Vol. 15, No. 6, 2019, pp. 3379-3387

[2] S. A. Reveliotis: Logical Control of Complex Resource Allocation Systems. Foundations and Trends in Systems and Control, Vol. 4, No. 1-2, 2017, pp. 1-223

[3] M. A. Lawley, S. A. Reveliotis: Deadlock Avoidance for Sequential Resource Allocation Systems: Hard and Easy Cases. The International Journal of Flexible Manufacturing Systems, Vol. 13, 2001, pp. 385-404

[4] S. A. Reveliotis: Coordinating Autonomy: Sequential Resource Allocation Systems for Automation. IEEE Robotics & Automation Magazine, Vol. 22, No. 2, 2015, pp. 77-94, DOI: 10.1109/MRA.2015.2401295

[5] M. W. Iordache, P. J. Antsaklis: Supervision Based on Place Invariants: A Survey, Discrete Event Dynamic Systems, Vol. 16, 2006, pp. 451-492

[6] Z. W. Li, M. C. Zhou: Deadlock Resolution in Automated Manufacturing Systems. A Novel Petri Net Approach. Springer-Verlag London Ltd., 2009

[7] J. C. Luo, Z. Q. Liu, M. C. Zhou: A Petri Net Based Deadlock Avoidance Policy for Flexible Manufacturing Systems with Assembly Operations and Multiple Resource Acquisition. IEEE Transactions on Industrial Informatics, Vol. 15, No. 6, 2019, pp. 3379-3387

[8] Z. Y. Ma, Z. W. Li, A. Giua: Petri net Controllers for Generalized Mutual Exclusion Constraints with Floor Operators. Automatica, Vol. 74, 2016, pp. 238-246

[9] M. M. Yan, Z. W. Li, N. Wei, M. Zhao: A Deadlock Prevention Policy for a Class of Petri Nets S^3PMR. Journal of Information Science and Engineering, Vol. 25, 2009, pp. 167-183

[10] G. Y. Liu, K. Barkaoui: A Survey of Siphons in Petri Nets. Information Sciences, Vol. 363, No. 1, 2016, pp. 198-220

[11] F. Tricas, J. Ezpeleta: Computing Minimal Siphons in Petri Net Models of Resource Allocation Systems: A Parallel Solution, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, Vol. 36, No. 3, May 2006, pp. 532-539, DOI: 10.1109/TSMCA.2005.855751

[12] A. Giua, M. Silva: Modeling, Analysis and Control of Discrete Event Systems: A Petri Net Perspective. In: 20th IFAC World Congress, Toulouse, France, IFAC PapersOnLine 50-1, 2017, pp. 1772-1783

[13] M. P. Fanti, B. Maione, S. Mascolo, B. Turchiano: Event-Bases Feedback Control for Deadlock Avoidance in Flexible Production Systems. IEEE Transaction on Robotics and Automation, Vol. 13, No. 3, 1997, pp. 347-363, DOI: 10.1109/70.585898

[14] Z. Y. Ma, Z. Li, A. Giua: Petri net Controllers for Generalized Mutual Exclusion Constraints with Floor Operators. Automatica, Vol. 74, 2016, pp. 238-246

[15] J. Desel, W. Reisig: Place/Transition Petri Nets. In: W. Reisig, G. Rozenberg (Eds.): Advances of Petri Nets, Lecture Notes in Computer Science, Vol. 1491, Springer, Heidelberg, 1998, pp. 122-173

[16] G. Y Liu, D. Y. Chao, F. Yu: Control Policy for a Subclass of Petri Nets without Reachability Analysis. IET Control Theory and Applications, Vol. 7, No. 8, 2013, pp. 1131-1141

[17] T. C. Row, W. M. Syu, Y. L. Pan, C. C. Wang: One Novel and Optimal Deadlock Recovery Policy for Flexible Manufacturing Systems Using Iterative Control Transitions Strategy. Mathematical Problems in Engineering, Vol. 2019, Article ID 4847072, 12 p.

[18] H. Yue, K. Y. Xing, H. S. Hu, W. M. Wu, H. Y. Su: PetriNet-Based Robust Supervisory Control of Automated Manufacturing Systems. Control Engineering Practice, Vol. 54, 2016, pp. 176-189

[19] F. Čapkovič: Modelling and Control of Discrete-Event Systems with Partial Non-Determinism Using Petri Nets, Acta Polytechnica Hungarica, Vol. 17, No. 4, 2020, pp. 47-66

[20] R. Davidrajuh: GPenSIM, General purpose Petri Net Simulator for Matlab Platform. Available: http://www.davidrajuh.net/gpensim/

Designing a New Data Encryption Algorithm Using a Genetic Code Method

Mustafa Zengin, Zafer Albayrak

Faculty of Engineering, Karabuk University, Computer Engineering Department, Baliklar Kayasi Mevkii 78050 Karabuk, Turkey,
mustafazengin@ogrenci.karabuk.edu.tr; zalbayrak@karabuk.edu.tr

Abstract: Today, the widespread use of information and communication tools along with the developing technology has facilitated access to information. These developments have revealed the importance of data security. Many encryption algorithms have been developed to ensure secure data transfer. In this article, we have developed a new Genetic Encryption Algorithm (GEA) inspired by the DNA structure. The GEA is compared to a DES (Standard Encryption Algorithm), an AES (Advanced Encryption Algorithm) and a RSA encryption algorithm. A short evaluation is made, presenting the results, along with tables and graphs.

Keywords: Cryptology; Genetics; Encryption; Algorithm; Performance

1 Introduction

The widespread use of computer technology has increased the importance of data security. With the effect of the Covid-19 epidemic, the use of internet and mobile devices has increased dramatically, especially studies in areas including, e-commerce, banking, finance, security and education are being carried out using the internet. A survey conducted during the Covid-19 progress, exploring the time that people spend on the Internet found the world average is 6 hours and 45 minutes. In a world where access to information is so easy, it has become an essential need to ensure data security. The secure transfer of data against attacks or threats, has become an important topic. The thought that data, that is needed to be kept confidential and correct, in the communications between computers, can be intercepted by unauthorized persons, is a big problem. The encryption of data is one of the simplest methods to ensure secure data exchange between computers. There are many encryption methods developed to ensure data security and protection of data [1]. These methods are explained in the subject of Cryptography. Cryptography is the process of making a message or data temporarily unreadable by passing it through various mathematical operations and converting this message to a normal readable state, upon reaching the desired target.

The cryptographic algorithms used today, are examined in two parts, symmetrically or asymmetrically, depending on the key structure [2]. In symmetric encryption algorithms, a single secret key is used for encryption and resolution of data. Using a single key while encrypting and resolving data creates a security problem. Because the key used in encryption is transmitted securely to the recipient and is used to resolve the same message, reveals the importance of key security. Symmetric encryption algorithms are faster and more efficient than asymmetric encryption algorithms [3]. However, they create security weaknesses because a single common key is used. DES, 3DES, AES are shown as the most widely used symmetric encryption algorithms [4]. The basic feature of symmetric encryption algorithms performs encryption by dividing the desired message into blocks and converting them into bits. For example, when the working principle of the DES algorithm is examined, it first divides the message into 64-bit blocks and then a 64-bit block back into 32-bit right and left bits. A 32- bit encrypted result is obtained by passing the 32-bit right bit through the f function with a 48-bit key bit. Then, the 32-bit left bit is passed through the f function with the same 48-bit key bit, and a 32-bit encrypted result is obtained. This f function produces a 32-bit result, using 32-bit data and a 48-bit key. This is the process of performing the action. This operation allows multiple results to be produced for the same bit. By extending with a 32-bit key, 48-bit data is provided. It splits the 48-bit data into groups, dividing the data into 8 blocks. Each block consists of a 6-bit segment. Each 6-bit piece has been reduced to 4 bits militarily this time. It consists of 8 blocks of 4 bits, that is, a total of 32 bits of data. A 64-bit block is obtained from 32-bit right and left messages that are encrypted [5]. With this method, all blocks are encrypted and form the working principle of the DES algorithm. Two public keys are used in asymmetric encryption algorithms. Using two different keys for encryption and resolution of data provides high security. However, compared to symmetric encryption algorithms, it is very slow and processing speed takes longer. The most widely used asymmetric encryption algorithm is the RSA algorithm. The main feature of asymmetric encryption algorithms is that they are easy to do, difficult to undo and time consuming because they operate with large prime numbers. For example, multiplying two numbers is easy, but finding their factors is difficult or takes time. Squaring a number is easy, but finding its square root is difficult. For this reason, asymmetric encryption algorithms are the most reliable encryption algorithms. The performance and success of cryptographic algorithms are determined according to the key size used in encryption, processing speed and the amount of memory used [6]. The “brute force” breaking times of the algorithms vary according to the key size used. The sample DES algorithm has a key structure of 56 bits. The brute force password cracking time is 2^{56} seconds. [7].

The aim of this study is to develop algorithms that perform better than the encryption algorithms used today in order to provide secure data transfer between large computer networks. GEA (Genetic Encryption Algorithm), which was developed as a result of the studies, has a symmetric encryption algorithm feature, so its processing speed is faster than asymmetric encryption algorithms [8].

Since the key size used is 128 bits, it is harder to crack a brute force password, it is more secure 2^{128} seconds. Because the brute force break time is The GEA encryption algorithm is faster than the DES, 3DES and RSA encryption algorithms, and the breaking time is more difficult than the DES and 3DES algorithms [9].

In the continuation of the study, in Section 2, cryptology and encryption algorithms, in Section 3, the studies in the literature, in Section 4, the structure of the genetic encryption algorithm developed inspired by the structure of DNA is explained. In Section 5, performance analysis of encryption algorithms is made and explained. The results of the study were evaluated in the last section.

2 Cryptology and Encryption Algorithms

Cryptology is the process of encrypting data and analyzing encrypted data. Encryption is the method and methods for encrypting data by performing the mathematical operations required to encrypt the data. Cryptanalysis is the science that covers the analysis of encrypted data. The purpose of encryption science is to provide data security to prevent data from falling into the hands of unwanted people. The purpose of password analysis is to analyze the data and convert existing passwords to their original state. Encryption, one of the basic concepts of cryptology, is the process of making plain text content unreadable. The main thing in encryption is to prevent data from being read by unauthorized people. The process of deciphering the password is the reverse of encryption and makes encrypted data meaningful, readable and understandable [10] [11].

When encryption methods are examined, it is seen that various techniques based on mathematical operations are used. While simple encryption techniques were used to provide data security in line with the opportunities provided by technological developments in history, modern encryption methods are used today. According to the characteristics of the keys used in modern encryption methods, they are divided into two as symmetric encryption algorithms and asymmetric encryption algorithms [12] [3].

2.1 Symmetric Encryption Algorithms

When we look at the structure and principles of symmetric encryption algorithms, a single key is used to encrypt data and recycle encrypted data and make it meaningfully readable. This key used is private.

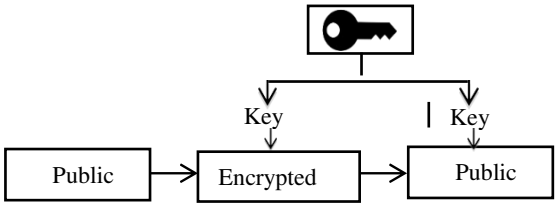


Figure 1
Symmetric Encryption Algorithms

A common key is used between people who provide encryption of data and analyze encrypted data. Therefore, in symmetric encryption algorithms, it is very important to securely transmit the key to the other party [8].

2.2 DES Encryption Algorithm

The standard data encryption algorithm uses the same key to encrypt data and recycle encrypted data. Therefore, the security of the key is very important. The working principle of the DES algorithm is shown in Figure 2.

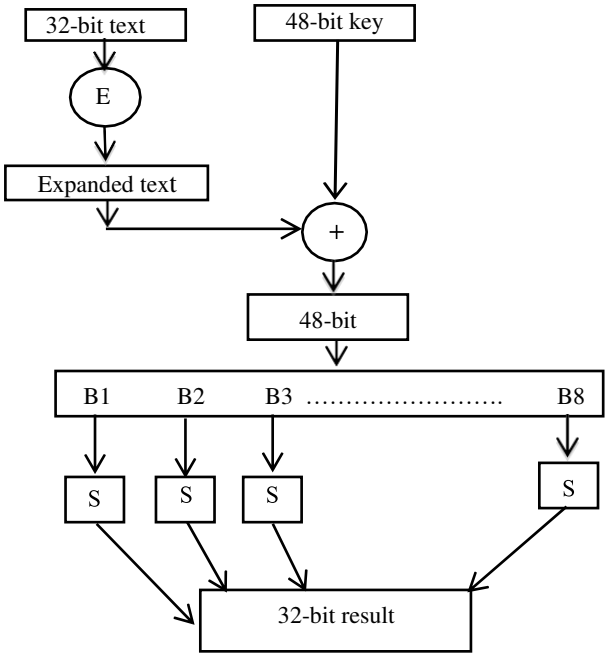


Figure 2
DES Encryption Algorithm

When the DES algorithm encrypts the data, it is processed in 64-bit blocks and the data is encrypted according to the symmetric encryption method with the help of a 56-bit key. Thus, 64-bit encrypted data is obtained [13]. The DES (Data Encryption Standard) algorithm is split into 64-bit data blocks encrypted with the same 56-bit key to restore data to 56-bit blocks. Thus, readable, meaningful data is obtained. As a result, data that seems meaningless by encrypting the data becomes meaningfully readable [4] [14].

2.3 AES Encryption Algorithm

The advanced data encryption algorithm uses the same key as the DES algorithm to encrypt and recycle encrypted data. For this reason, the security of the key to be used in encryption is very important. The working principle of the AES algorithm is shown in Figure 3.

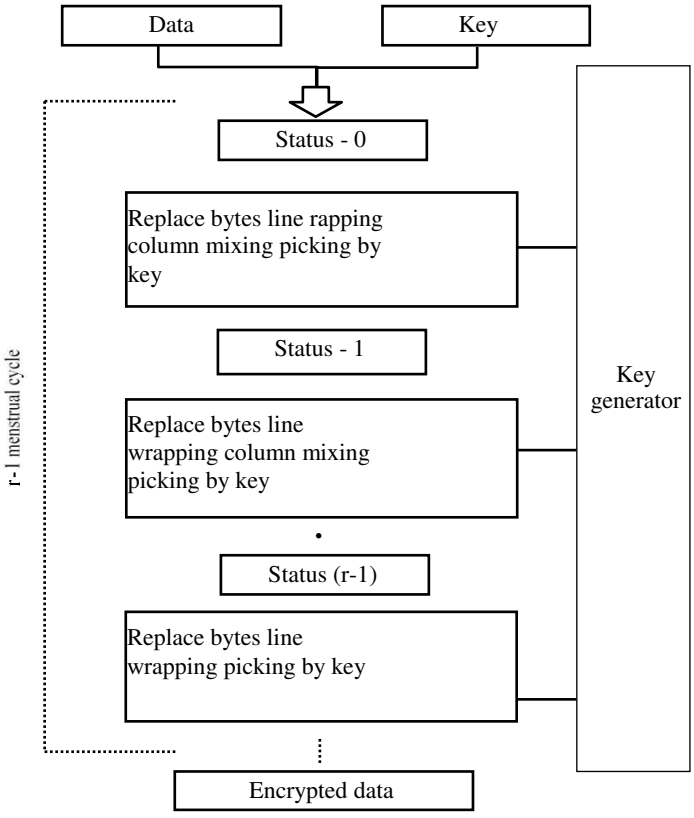


Figure 3
AES Encryption Algorithm

The AES symmetric encryption algorithm consists of a return transformation block and a key generation block. Because the Advanced Encryption Standard is a repetitive algorithm, if 128-bit keys are used for encryption or decryption, it results in 10 repeats, 12 repeats in 192-bit keys, and 14 repeats in 256-bit keys [15]. The flexible nature of the AES encryption algorithm does not adversely affect processing speed and performance, even when using a different key [16].

2.4 Asymmetric Encryption Algorithms

Two different keys are used in these encryption algorithms. A different key is used when encrypting data and deciphering encrypted data. The working principle of asymmetric encryption algorithms is shown in Figure 4. Two different keys are used in asymmetric encryption algorithms. It is not hidden like the key in symmetric encryption algorithms. In addition, a single secret key is used in symmetric encryption algorithms, while two keys are used in asymmetric encryption algorithms. One of the keys is the public key. The other key is hidden. The public key is used to encrypt the data, while the secret key is used to analyze the encrypted data [17].

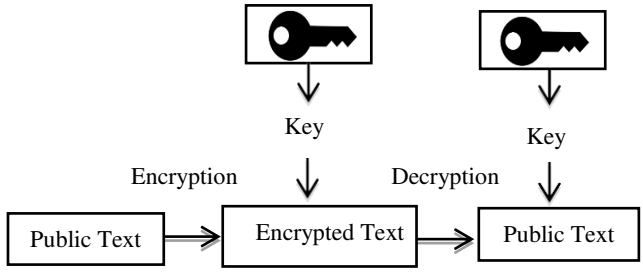


Figure 4
Asymmetric Encryption Algorithms

2.5 RSA Encryption Algorithms

The key used to encrypt data in the RSA encryption algorithm is public. The same key is not used to decrypt encrypted data. The working principle of the RSA encryption algorithm is shown in Figure 5. The security of the RSA algorithm is based on the algorithmic difficulties of factoring the numbers. This asymmetric encryption algorithm is used as a public key, along with another value selected by the product of two large prime numbers. Prime multipliers are stored. The data can be encrypted using the public key, but if the public key is large enough, the encrypted message can only be deciphered if the prime number multiplier is known [4].

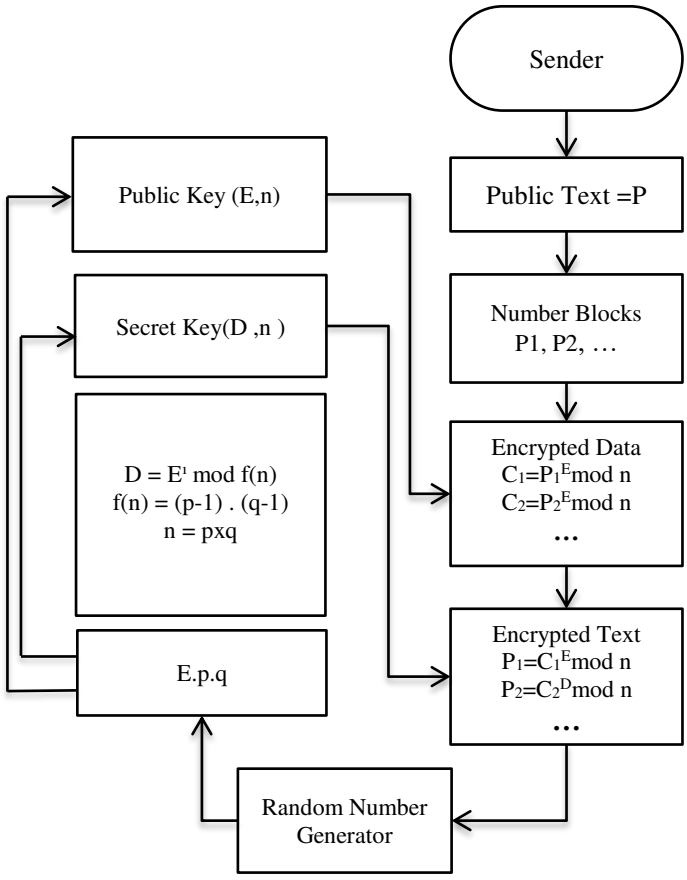


Figure 5
RSA Encryption Algorithm

3 Related Literature Studies

As a literature research for performance analysis of encryption algorithms many articles have reviewed and detailed. In [18], a study was conducted on cryptography algorithms to perform performance and security analysis of simple encryption algorithms. While encrypting on different image files, data distribution, pixel count comparison, encryption time and encryption quality analyzes were performed. They concluded that S-AES and LBlock algorithms provide fast and sufficient security using less resource. In [19], information was given about the methods used in classical encryption methods and methods used in modern encryption. In the encryptions made with the help of the key used in modern encryption methods, the working time of the algorithms, the processor and memory usage features of the algorithms were examined. This work was done only on pixels and images.

In [20], parameters such as the accuracy, efficiency and key exchange of a preferred algorithm for BLOWFISH, IDEA, CAST-128, RC6, DES, 3DES, AES and RSA encryption were analyzed. It is emphasized that it can be provided by applying to multiple algorithms to create efficient encryption systems.

Thakur et al, Among the symmetric encryption algorithms, the most commonly used DES, AES and Blowfish symmetric encryption algorithms were examined and performance analyzes were compared in terms of speed, block size and key size. As a result of the java simulation program used, Blowfish showed that the encryption algorithm has better performance. [21]

In [22], DES, AES, 3DES, RC2, Blowfish, IDEA, Twofish, TEA, symmetric encryption algorithms and RSA asymmetric encryption algorithm were used to ensure data security. While these encryption algorithms encrypt the data, time, memory and processor usage performance criteria are compared. As a result of the studies, the DES algorithm performed better than the AES algorithm in small data sizes. The DES algorithm did not perform well on large data sizes compared to the AES algorithm.

In [23] symmetric and asymmetric encryption algorithms are examined. Key sizes were analyzed during encryption or decryption. They examined the factors that affect the performance of data encryption algorithms.

Matching of organic bases in Deoxyribbo Nucleic Acid (DNA) structure was investigated. As a result of this research, the matching of the bases inspired us to develop a symmetric encryption algorithm with 128-bit random key structure [24].

4 Designing a New Data Encryption Algorithm Using Genetic Code Method (GEA)

The DNA structure, in which the biological properties of living things are carried, forms the basis of our algorithm. Structure of DNA (Deoxyribbo Nucleic Acid) There are organic bases such as Adenine "A", Guanine "G", Cytosine "C" and Thymine "T". A different encryption algorithm has been developed based on the relationships of these organic bases with each other [25].

In this study, While ASCII coding of Adenine, Thymine, Guanine and Cytosine bases in the structure of DNA was done, the "Watson Crick's" model was used. This method is linked by hydrogen bonds between Adenine and Thymine, Guanine and Cytosine base in DNA. The reciprocal substitution of these bases introduces complexity in terms of encryption. This complexity constitutes the security of the encryption process. The data to be encrypted and the ASCII values of the keyword to be used were calculated. These ASCII values, which are calculated as a decimal number system, are converted into genetic code. The ASCII value of each character is converted into a quad number system. Numerical values are created. (When we

convert the data in the decimal number system to the system of four numbers, the remaining values are 0, 1, 2, and 3.)

The main reason for converting our ASCII values to four number systems is to match four organic bases. This match is shown in Figure 6. The mapping of DNA bases on the left. By assigning a number value to each of the middle DNA bases, which base matches which number matches. The figure on the right shows how DNA bases match the number values we assign.

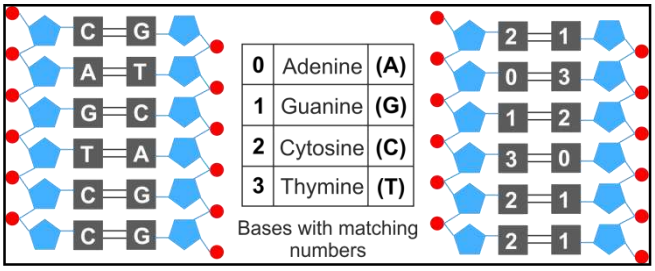


Figure 6
Coding of DNA Organic Bases

In this way, our numbers turn into organic bases that match the corresponding organic bases, as in DNA matching. (There is a match between Adenine and Thymine, Guanine and Cytosine, and vice versa.) This 3 - 0, 2 - 1. Thus, the quad number system creates a new DNA chain. The keyword match used in encryption with this DNA match is aggregated in a quadruple system. A single DNA chain is created. In the quad number system, these newly created numbers are converted to their equivalents in the decimal system. Finally, character values in the ASCII number table in the decimal system are obtained. Although these values seem meaningless and complex, they show encrypted data. By recycling these encrypted ASCII values, the encrypted data is converted to the original state, that is, by reversing these processes to decode the encrypted data [26] [27]. Among the biggest disadvantages of symmetric encryption algorithms is the use of a single key when encrypting and decrypting data, and the small key size. In this study, greater usability of the key size is provided.

5 Performance Analysis of Symmetric and Asymmetric Encryption Algorithms

One of the reliable methods used in performance evaluation of complex encryption algorithms is experimental analysis method [26] [27]. For this reason, performance analyses of symmetric and asymmetric encryption algorithms were done using experimental analysis method. DES, AES and RSA encryption algorithms were compared to see the performance values of the encryption algorithm (GEA)

application with the developed genetic code method. The performance of GEA, DES, AES and RSA algorithms of the data to be encrypted was calculated by using experimental measurements in accordance with the processor (CPU) and memory (RAM) usage values during this process (encryption and decryption). In this analysis method, the processing time was evaluated in minutes, the memory megabyte used and the processor % used. This experiment was calculated according to the encryption and decryption of data packets of 58 Bytes, 102 Kilobytes, 1 Megabyte and 5 Megabytes.

5.1 Encryption and Decryption Analysis of 58 Bytes of Data

The process time, processor usage, and memory usage values obtained during the encryption and decryption of data with 58 bytes of character length were examined. Table 1 shows the performance values of DES, AES, GEA and RSA encryption algorithms when 58 bytes of data are encrypted.

Table 1
58 Byte Data Encryption

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	2	30	8
	AES	3	32	8
	RSA	8	29	8
	GEA	2	18	8

When the graphic in Figure 7 is examined, it is seen that the processor values used when encrypting 58 bytes of data are the same.

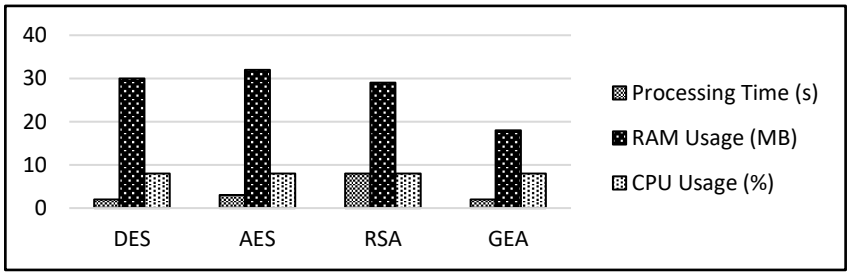


Figure 7
Byte Data Encryption

Although the processing time value of the GEA algorithm is the same as the DES algorithm, memory usage values perform better than other DES, AES and RSA algorithms. The reason for this is the switch structure used and the working principle used. While the GEA uses a 128-bit key structure, the DES algorithm uses a 56-bit key structure. Although this affects the processing speed, it increases the memory usage of the DES algorithm.

Table 2
Decryption of 58 Bytes of Data

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	1	26	8
	AES	1	28	8
	RSA	6	24	8
	GEA	1	14	8

Table 2 shows the performance values of DES, AES, GEA and RSA encryption algorithms for deciphering data of 58 bytes. While analyzing the 58 byte encrypted data in the analysis of the graph in Figure 8, the processing time and processor usage performance of DES, AES and GEA symmetric encryption algorithms are better than RSA, which is the asymmetric encryption algorithm. Symmetric encryption algorithms have the same values.

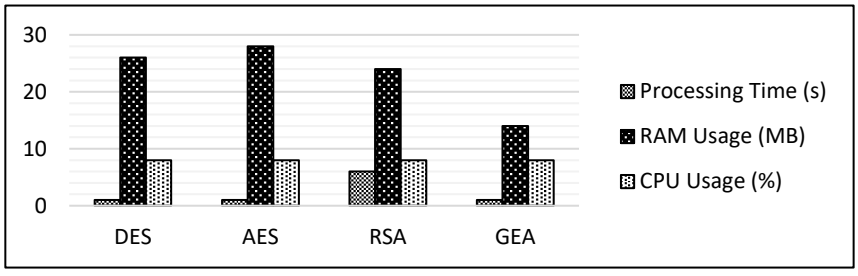


Figure 8
Decryption of 58 Byte Data

In general, the memory usage performance of GEA algorithm is seen to be better than that of DES, AES and RSA algorithms. This is due to the use of 128-bit key and working principle.

5.2 102 Encryption and Decryption Analysis of Kilobyte Data

The values for the processing time, processor usage and memory usage obtained during the encryption and decryption of data with 102 kilobyte character length were examined.

Table 3 shows the performance values of DES, AES and GEA symmetric encryption algorithms while encrypting 102 kilobytes of data. RSA asymmetric encryption algorithm cannot make big data encryptions. Its algorithm is not suitable for this. It usually works on large prime numbers. Therefore, it will not be included in the above and subsequent data analysis.

Table 3
102 Kilobyte Data Encryption

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	171	35	24
	AES	173	42	26
	GEA	169	32	22

When the graphic in Figure 9 is analyzed, it is seen that GEA algorithm processing time, memory and CPU usage performance are better than other DES and AES algorithms. The reason for the good performance of the GEA symmetric encryption algorithm is due to the key size used and the working principle of the algorithm.

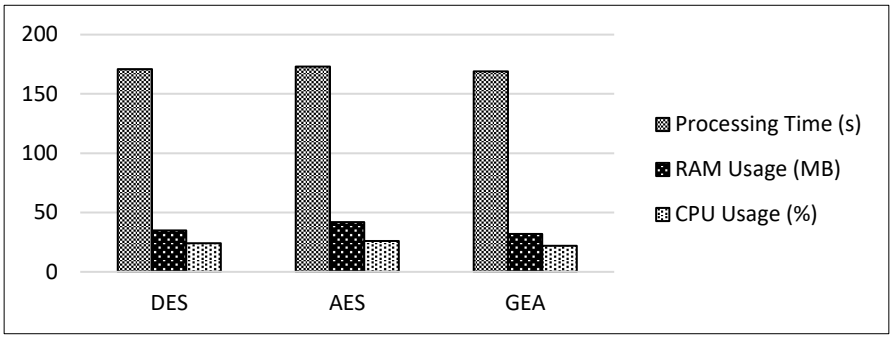


Figure 9
102 Kilobyte Data Encryptions

Table 4 shows the performance values of DES, AES and GEA symmetric encryption algorithms in the process of deciphering 102 Kilobyte data.

Table 4
Decryption Table of 102 Kilobyte Data

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	154	28	24
	AES	168	36	26
	GEA	148	26	22

When the graphic in Figure 10 is examined, it is seen that the processing time, memory and processor usage values of the GEA algorithm perform better than the DES and AES algorithms. This is just because of the key size and working principle used when encrypting data.

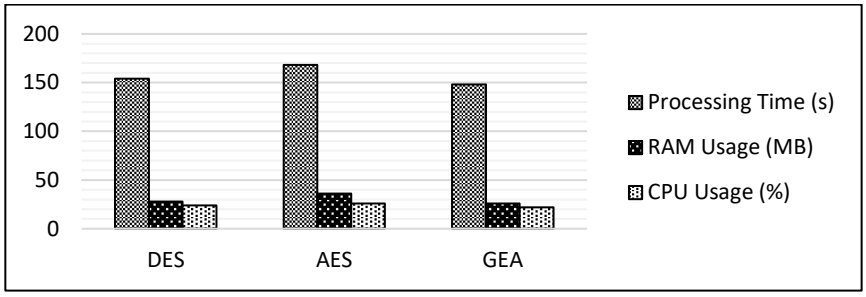


Figure 10
Decryption of 102 Kilobytes Data

5.3 Encryption and Decryption Analysis of 1 Megabyte Data

The processing time, processor usage and memory usage values obtained during the encryption and decryption of data with a character length of 1 Megabyte were examined.

Table 5
1 Megabyte Data Encryption Table

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	854	168	36
	AES	751	164	34
	GEA	853	166	35

Table 5 shows the performance values of DES, AES and GEA symmetric encryption algorithms for encrypting 1 Megabyte data.

When the graphic in Figure 11 is examined, it is seen that the AES encryption algorithm, memory and processor usage performance is better than DES and GEA algorithms. The reason that the AES algorithm performs better is due to the structure and working principle of the AES algorithm. The AES algorithm is that it has a flexible structure. Processing speed and performance do not change even if 128-bit, 192-bit or 256-bit keys of different sizes are used.

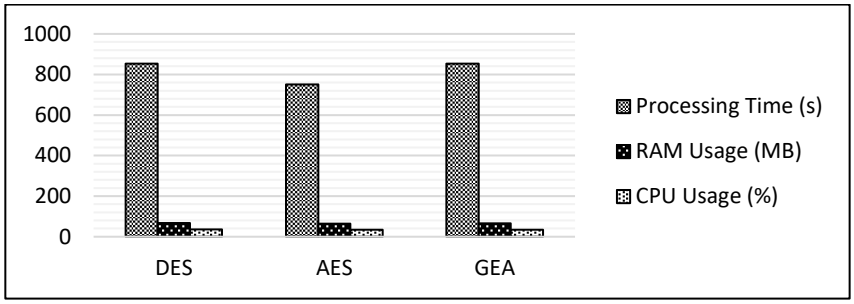


Figure 11
1 Megabyte Data Encryption

Table 6 shows the performance values of DES, AES and GEA symmetric encryption algorithms for password decoding of 1 Megabyte data.

Table 6
Decryption Table of 1 Megabyte Data

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	648	142	34
	AES	587	158	32
	GEA	643	160	33

When the graphic in Figure 12 is analyzed, it is seen that the performance of AES encryption algorithm, memory and processor usage is better than DES and GEA algorithms. The reason for the good performance of the AES symmetric encryption algorithm is the structure and working principle of the AES encryption algorithm.

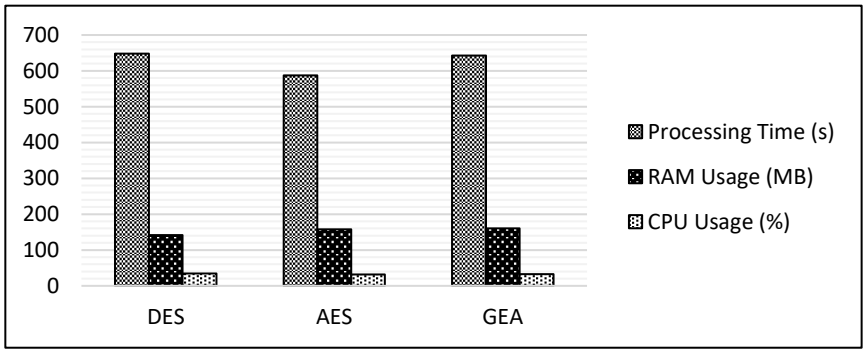


Figure 12
Decryption of 1 Megabyte Data

5.4 Encryption and Decryption Analysis of 5 Megabyte Data

The processing time, processor usage and memory usage values obtained during encryption and decryption of 5 Megabyte character length data were examined. Table 7 shows the performance values of DES, AES and GEA symmetric encryption algorithms for encryption of 5 Megabyte data.

Table 7
5 Megabyte Data Encryption

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	868	168	44
	AES	856	162	40
	GEA	886	174	54

When the graphic in Figure 13 is examined, it is seen that the performance values of the AES encryption algorithm are better, while the DES and GEA algorithm processing time, memory and processor usage performance values are very close to each other.

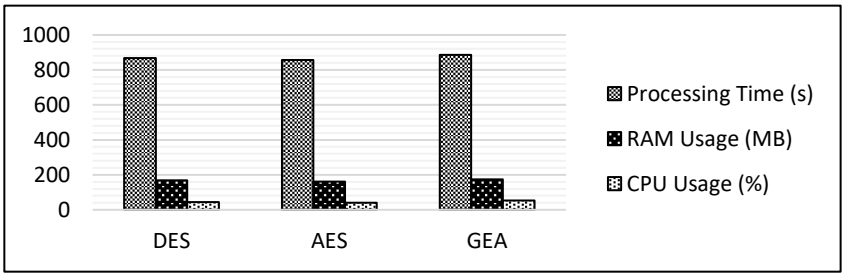


Figure 13
5 Megabyte Data Encryption

The reason for the good performance of the AES encryption algorithm is that it is more successful in terms of working principle in encrypting big data. Table 8 shows the performance values of DES, AES and GEA symmetric encryption algorithms for deciphering 5 Megabyte encrypted data.

Table 8
Decryption Table of 5 Megabyte Data

	Encryption Algorithms	Processing Time (s)	RAM Usage (MB)	CPU Usage (%)
Tested Computer	DES	648	142	34
	AES	587	138	32
	GEA	643	140	33

When the graphic in Figure 14 is examined, it is seen that the performance values of the AES encryption algorithm are better than the performance values of the DES and GEA encryption algorithms in the analysis of 5 Megabyte encrypted data. The reason for the better performance of the AES encryption algorithm is that the algorithm has its own unique working principle.

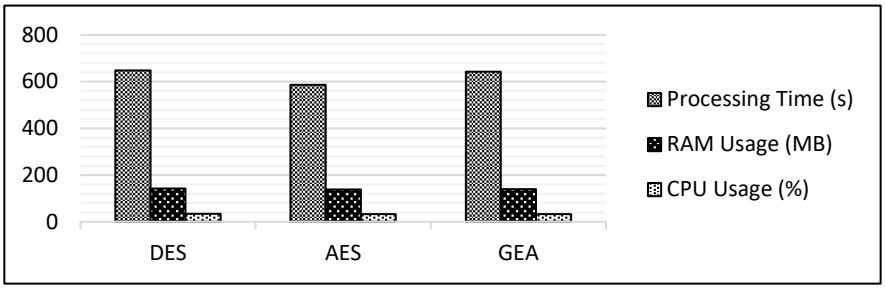


Figure 14
Decryption Graph of 5 Megabyte Data

Conclusions

Symmetric encryption algorithms performed better than asymmetric encryption algorithms, in encrypting data and analyzing encrypted data. Depending on the size of the key used for encryption and decryption, brute force breakage times vary. Since the DES algorithm uses a 56-bit key, the brute force break time is 2^{56} seconds. AES and GEA algorithms use 128-bit keys; the brute force break time is 2^{128} seconds. Genetic encryption algorithm developed to encrypt and decrypt small size data. DES performed better than AES and RSA encryption algorithms. The performance of the GEA algorithm is due to the amount of memory used for data encryption and recycling of encrypted data, processing speed, key size used and working principle. The AES encryption algorithm has been shown to be more successful in encrypting and decrypting large data. Since the RSA encryption algorithm does not tend to encrypt big data, it performs encryption based on the multiplication of large prime numbers, based on mathematical operations.

References

[1] C. C. Chang, M. S. Hwang, and T. S. Chen, "A new encryption algorithm for image cryptosystems," *J. Syst. Softw.*, 2001, doi: 10.1016/S0164-1212(01)00029-2

[2] K. Raeburn and MIT, "RFC 3962 - Advanced Encryption Standard (AES) Encryption for Kerberos 5," 2005

[3] W. M. H. Company, "Modern Cryptography: Theory and Practice," *Theory Pract.* 2003

[4] R. Bhanot and R. Hans, "A review and comparative analysis of various encryption algorithms," *Int. J. Secur. its Appl.*, 2015, doi: 10.14257/ijsia.2015.9.4.27

[5] R. Tripathi and S. Agrawal, "Comparative Study of Symmetric and Asymmetric Cryptography Techniques," *Int. J. Adv. Found. Res. Comput.*, 2014

[6] M. Faheem, S. Jamel, A. Hassan, Z. A., N. Shafinaz, and M. Mat, "A Survey on the Cryptographic Encryption Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, 2017, doi: 10.14569/ijacsa.2017.081141

[7] T. Nie and T. Zhang, "A study of DES and blowfish encryption algorithm," 2009, doi: 10.1109/TENCON.2009.5396115

[8] E. Islam and S. Azad, "Data encryption standard," in *Practical Cryptography: Algorithms and Implementations Using C++*, 2014

[9] P. P. Churi, "Performance analysis of data encryption algorithm," *Int. J. Recent Technol. Eng.*, 2019, doi: 10.35940/ijrte.C5775.098319

[10] P. Kumar and S. B. Rana, "Development of modified AES algorithm for data security," *Optik (Stuttg).*, 2016, doi: 10.1016/j.ijleo.2015.11.188

[11] A. Mousa, "Data encryption performance based on Blowfish," 2005, doi: 10.1109/elmar.2005.193660

[12] A. G. Radwan, S. H. AbdElHaleem, and S. K. Abd-El-Hafiz, "Symmetric encryption algorithms using chaotic and non-chaotic generators: A review," *Journal of Advanced Research*. 2016, doi: 10.1016/j.jare.2015.07.002

[13] J. Grabbe, "The DES algorithm illustrated," *Laissez Faire City Times*, 1992

[14] M. Prerna and S. Abhishek, "A Study of Encryption Algorithms AES, DES and RSA for Security," *Glob. J. Comput. Sci. Technol. Network, Web Secur.*, 2013

[15] M. J. Dworkin, "FIPS 197, Advanced Encryption Standard (AES)," *Netw. Secur. Natl. Inst. Stand. Technol.*, 2001

[16] S. D. Rihan, A. Khalid, and S. Eldin F. Osman, "A Performance Comparison of Encryption Algorithms AES and DES," *Intetnational J. Eng. Res. Technol.*, 2015

[17] S. Ahmad, K. M. R. Alam, H. Rahman, and S. Tamura, "A comparison between symmetric and asymmetric key encryption algorithm based decryption mixnets," 2015, doi: 10.1109/NSysS.2015.7043532

[18] Ü. Çavuşoğlu and H. Al-Sanabani, "The Performance Comparison of Lightweight Encryption Algorithms," *Sak. Univ. J. Comput. Inf. Sci.*, 2019, doi: 10.35377/saucis.02.03.648493

[19] G. F. S. M. Asfiya Shireen Shaikh Mukhtar, "An Introduction of Advanced Encryption Algorithm: A Preview," *Int. J. Sci. Res.*, 2014

[20] K. Aggarwal, J. Kaur Saini, and H. K. Verma, "Performance Evaluation of RC6, Blowfish, DES, IDEA, CAST-128 Block Ciphers," *Int. J. Comput.*

Appl., 2013, doi: 10.5120/11749-7244

[21] J. Raigoza and K. Jituri, "Evaluating Performance of Symmetric Encryption Algorithms," 2017, doi: 10.1109/CSCI.2016.0258

[22] M. Mathur, "Comparison between DES, 3DES, RC2, RC6, BLOWFISH and AES," *Proc. Natl. Conf. New Horizons IT*, 2013

[23] A. L. Jeeva, D. V Palanisamy, and K. Kanagaram, "Comparative analysis of performance efficiency and security measures of some encryption algorithms," *Int. J. Eng. Res. Appl. ISSN*, 2012

[24] A. Majumder, A. Majumdar, T. Podder, N. Kar, and M. Sharma, "Secure data communication and cryptography based on DNA based message encoding," 2015, doi: 10.1109/ICACCCT.2014.7019464

[25] B. Guttman, "DNA Structure," in *Brenner's Encyclopedia of Genetics: Second Edition*, 2013

[26] Şatır Esra Talu Furkan, "DES Algoritmasının DNA Modellenmesi ile Performans Analizi," International Marmara Sciences Congress, 2020

[27] Kovalchuk, A., Izonin, I., Strauss, C., Podavalkina, M., Lotoshynska, N., & Kustra, N. (2019) Image Encryption and Decryption Schemes Using Linear and Quadratic Fractal Algorithms and Their Systems. In *DCSMart* (pp. 139-150)

[28] D. S. Abd Elminaam, H. M. A. Kader, and M. M. Hadhoud, "Evaluating the performance of symmetric encryption algorithms," *Int. J. Netw. Secur.*, 2010

[29] D. S. Abdul, H. M. Abdul Kader, and M. M. Hadhoud, "Performance evaluation of symmetric encryption algorithms," 2009