ACADEMY OF SCIENCES OF MOLDOVA
INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE

Title of manuscript
U.D.C: 519. 95

ALBU VEACESLAV

HUMAN ACTIONS RECOGNITION WITH MODULAR NEURAL NETWORKS

SPECIALTY: 122.03

MODELING, MATHEMATICAL METHODS, SOFTWARE

Abstract of the Ph. D. Thesis in Computer Science

Chisinau, 2016

This thesis has been elaborated with the assistance of the “Programming Systems” Laboratory of the Institute of Mathematics and Computer Science, Academy of Sciences of Moldova.

Scientific Adviser: **COJOCARU Svetlana**, Doctor in Habilitation in Computer Science, Prof.

Official Reviewers:
VAGHIN Vadim, Doctor of Technical Sciences, Prof., Moscow Power Engineering Institute.
CĂPĂȚÂNĂ Gheorghe, Doctor in Computer Science, Prof., Moldova State University.

Members of the Specialized Scientific Council:

GAINDRIC Constantin, ***President***, Dr. Hab. in Computer Science, Professor, Corresponding Member of the Academy of Sciences of Moldova, Institute of Mathematics and Comp. Science.
CIUBOTARU Constantin, ***Scientific Secretary***, Dr. in Computer Science, Associate Professor, Institute of Mathematics and Computer Science, Academy of Sciences of Moldova.
COSTAŞ Ilie, Dr. Hab. Computer Science, Professor, Academy of Economic Studies, Chişinău.
GUȚULEAC Emilian, Dr. Hab. Computer Science, Professor, Technical University of Moldova, Chişinău.
AVERKIN Alexei, Doctor of Technical Sciences, Associate Professor, Computing Center of Academy of Sciences of Russia, Moscow.
BURȚEVA Liudmila, Dr. in Computer Science, Associate Professor, Institute of Mathematics and Computer Science, Academy of Sciences of Moldova.
ŢIȚCHIEV Inga, Dr. in Computer Science, Associate Professor, Institute of Mathematics and Computer Science, Academy of Sciences of Moldova.

The Ph. D. thesis shall be presented on November, __9__, 2016, 15.00, at the session of the Specialized Scientific Council DH 01.122.03 – 03, at the Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, str. Academiei 5, Chişinău, MD-2028, Republic of Moldova.

The Ph. D. thesis and its abstract can be accessed at the “Andrei Lupan” Central Library of the Academy of Sciences of Moldova and on the Web page of C.N.A.A. (www.cnaa.md).

The abstract of the Ph.D. thesis has been sent at October, ____2016.

Scientific Secretary of the Specialized Scientific Council:
CIUBOTARU Constantin,
Dr. Comp. Sci., Assoc. Prof. ________________

Scientific Advisor:
COJOCARU Svetlana,
Dr. Hab. in Computer Science, Prof. ________________

Author:
Veaceslav Albu

1. CONCEPTUAL PERSPECTIVES ON THE THESIS

The actuality and significance of the emotion and gesture recognition problem

Humans possess a remarkable ability to recognize objects very accurately by simply looking at them. However, when we study the underlying neuronal processes, they appear to be extremely complicated: the recognition process in primate visual cortex involves many areas and relatively high processing complexity.

The artificial system, which will try to mimic all the functions of natural recognition system will either be too complicated to construct or will acquire the computational complexity which is hard to attain. Therefore, the artificial recognition system usually simplifies the matters and in this research, we also model only the general functional principles of the neuronal organization of visual areas. However, we will try to achieve neurophysiological plausibility and maintain as high level of detail as possible. Moreover, the additional complexity is added up by the requirement to recognize the moving object in real time, i.e. to recognize not only static images, but a real-time video flow, which adds the temporal component to the recognition process.

Our answer to the complex problem of the emotion and gesture recognition problem is to propose an artificial neural network (ANN) for classification of human gestures and emotions, obtained from infrared cameras. The output from the cameras serves as an input into the proposed network and obtain the classifications of person’s reactions into typical vs. non-typical for an interaction with a certain type of environment. The proposed ANN can serve as a robust tool for classification of emotions and gestures of a human subject into typical vs. non-typical for a certain kind of interaction in real-time by utilizing the state-of-the art machine learning algorithms, originated from biologically plausible neural network (NN) architectures.

State-of-the-Art and problems in the emotion and gesture recognition field

Emotion classification. In the field of emotion recognition, there are three main problems, that require clarification. The first difficult conceptual problem, underlined by many researchers is the concept emotion. Among the questions that arise here, one significant is how to distinguish emotion differ from other facets of human experience? The lack of a clear definition of emotion has caused much difficulty for those trying to study the face and emotion.

We will provide the definitions from the classic research in the field of emotion recognition and classification and some of the contemporary researchers to choose the best definition that can serve our purposes. Another difficult conceptual problem is specifying the emotions accurately. How do we know whether information provided by the face is accurate? Is there someone criterion to determine what emotion has actually been experienced? In the experimental section, we have conducted a series of psychological experiments with human subjects in order to define the exact emotion of the person from personal judgments and from the comments of human observers.

These two problems are regarded independently with the second and most important one: how to recognize the emotion and action in real time from a video flow? To solve it, we use the insights from computational neuroscience to build our model. In this thesis, we will understand emotions as guides or biases to behaviours and decision making, which are possible to measure through measuring the visible facial features. There are a number of models of emotions developed for different purposes like formalization, computation or understanding. All models of emotions can be classified into discrete and continuous. Discrete models work with limited sets of emotions.

The best know and the most widely used discrete model of emotions was developed by Paul Ekman [1]. He developed his model over years and ended up with six basic emotions: anger, disgust, fear, happiness, sadness, and surprise.

Most of the works in this field are reduced to recognition of five emotions (i.e. disgust, fear, joy, surprise, sadness, anger), following Ekman and Friesen [2]. On the other hand, these emotions in their pure expression are rarely met in real life, a person’s emotional state being characterized by a spectrum of expressions. Typically, emotions are manifested trough some minor actions that alter facial features, such as lip corners raised or lowered in a state of joy or sadness.

Therefore, in the proposed work we use the data from our own psychological experiments to define facial expressions [3]. Facial expressions are accessed two-ways: from the personal reference of a human subject and from the judgement of human observer. Nevertheless, we use the labels, proposed by Ekman in his work, excluding the ones we have never observed throughout the experiments.

Action classification. Human action recognition is the process of labelling image sequences with action labels. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. The task is challenging due to variations in motion performance, recording settings and inter-personal differences. A number of attempts were done to approach real-time video classification with neural networks.

One of the recent breakthrough in this field belongs to Karpathy et al.[4]: they have studied the performance of CNNs in large-scale video classification. They proved that CNN architectures are capable of learning features from weakly-labelled data that is better than feature- based methods in performance and that these benefits are surprisingly robust to details of the connectivity of the architectures in time. Also, they have suggested that more careful treatment of camera motion may be necessary (for example by extracting features in the local coordinate system of a tracked point). In our system, this problem is non-existent, since the camera is fixed and the user is usually located at the same position in front of infrared camera. Other problems are addressed accordingly with application of the deep CNNs for input video classification. Also, we use the output from infrared cameras (depth maps) as an input to our system, which simplifies the recognition process and makes it more accurate.

The main purpose of this thesis

The main purpose of the presented research is to develop a tool for classification of human reactions (including both emotions and actions) into typical and non-typical in real time in a certain environment. This tool provides statistical observations and measurements of human emotional states during an interaction session with a software product (implemented in a slightly augmented hardware platform). Using computer vision and machine learning algorithms, emotions are recorded, recognized, and analyzed to give statistical feedback of the overall emotions of a number of targets within a certain time frame.

Similarly, we classify the actions of human subjects, which a user can perform during the interaction with a piece of software/hardware complex and provide a classification of his actions. The feedback, produced by the proposed system, can provide important measures for user response to a chosen system.

An application example of this research is a camera system embedded in a machine that is used frequently, such as an ATM. We use camera recordings to capture the emotional state of customers (happy, sad, neutral, etc.) and build a database of users and recorded emotions to be analyzed later. For the purposes of the study, we have developed and tested a hardware complex, which we use to conduct psychological experiments.

Objectives of the work

The main research objectives of the presented work could be formulated as following:

1. To develop a tool for classification of emotions and actions of a human subject into two groups (typical vs. non-typical) for a certain kind of interaction. We propose neural network architecture for classification of human gestures and emotions, obtained from infrared cameras.

The output from the cameras serves as an input into the proposed network, which classify human's reactions into typical vs. non-typical during an interaction with a certain type of environment. Here, the term 'reaction' refers to the combination of emotions and body movements, preformed by a human subject.

For academic purposes, we have chosen a very limited number of emotional states and behavioural patterns by studying only type of such standard interaction: the interaction of a user with typical ATM equipment, since it provides us with very distinctive patterns of 'typical' and 'non-typical' facial expressions. During this study, we observed the behaviour of human subjects during standard interaction with the ATM versus non-standard interaction.

Automated analysis of these behaviours with the machine learning techniques allowed us to train a complex convolutional neural network (CNN) to make an inference about behaviour of a user by classification both body movements and facial features. Such a feedback can provide important measures for user response during an interaction with any chosen system with a limited number of gestures involved. We use infrared cameras to automatically detect features and the movements of the limbs in order to classify user behaviour into typical or untypical for the kind of task he is performing.

The aim of current paper is to analyse the person's actions during the interaction with a user interface and implement the algorithm, which will be able to classify the human behaviour from infrared sensor input (normal vs. abnormal) in real time.

2. Among all the state-of-the art approaches, that are commonly used for both gesture and emotion classification, to choose one, that will be robust, high-performing and allow recognition of selected features. We develop and test two types of the algorithms, which could be applied in such a system and compare the results of these studies.

The reason for us to choose two types of neural networks is the condition that we analyse two types of features (facial features and gestures) simultaneously, which requires substantial computational costs. We suggest using deep neural network in combination with radial basis function network (the details would be provided in chapter two). However, second type of neural network could be used alone for this type of task.

3. To conduct behavioural experiments in order to evaluate how effectively the proposed system can detect normal vs. abnormal behaviour of a customer during interaction with ATM and make a conclusion about the applicability of the proposed system to industrial/commercial purposes.

Methodology of research

Throughout the study, we will introduce *two main research methods* we utilize to build the software. Both of the methods originate from neural network theory, therefore we introduce the theory of neural networks in detail in chapter two. We provide the detailed mathematical notation for every part of the model, including the learning algorithm. The learning algorithms we use for the two parts of the system are very similar, though differ in some details. We use some concepts from the field of machine learning, since it constituted the large part of this study.

The novelty and scientific originality of this thesis consists in a novel modular neural network architecture, constituted from two separate parts and combine the results to introduce the classification of the infrared sensor inputs, which is the first system of this kind, being applied both to emotion and human action recognition.

More exactly, we propose a combination of most recent biometric techniques with the NN approach for real-time emotion and behavioural analysis. Emotion and action recognition techniques have been presented separately in multiple studies during the past 5 years. However, the holistic approach has not been presented so far. Moreover, we present our algorithm in a framework of application to a particular task.

Theoretical significance

Our research solutions provide ground for solving of following problems: formulation of the tool's architecture for robust classification of emotions and gestures of a human subject into typical vs. non-typical; the substantiation of the possibility and efficiency of using deep learning in an integrated approach for the detection of expression of the whole body in real time**.**

From this point of view, our contribution is two-fold: we offer a novel neural network architecture, constituted from two separate parts and combining the results to introduce the classification of the infrared sensor inputs. To our knowledge, it is the first system of this kind, being applied to human action and emotion recognition. Parts of this system (like video processing, emotion recognition with convolutional networks etc.) were implemented before, but the whole realization is new. Moreover, the existing algorithms were modified to large extent (e.g. conventional SOM algorithm) for the purposes of this study.

Applied value of the work

The applications of this approach are possible in the variety of fields, including security systems, surveillance camera systems, biometrics etc.

The important scientific problem solved in this study is elaboration of a multimodal method for classification of human reactions (joining emotions and actions) into typical and non-typical in a certain environment, that ensures an effective functioning of systems destined to human actions monitoring in real time.

Main scientific results promoted for defence

The overall system performance, based on the experimental results, can be summarized as following:

1) The architecture of basic module of the network comprised of the self-organized map (SOM) of functional radial-basis function (RBF) modules is proposed, its mathematical foundation is presented. The proposed approach is new from the point of view of system architecture and the implementation of learning algorithm and, as we are aware, this architecture has never been applied to the task of emotion recognition.

2) The possibility of adapting the convolutional neural network architecture to a new type of input processing (infrared) has been demonstrated. It was shown that such kind of architecture is able to solve our task (action processing) in real time.

3) The developed NN model is able to recognize and classify emotions and body movements into two types (typical and non-typical) and facial expression has the accuracy of 8% and 14% error rate, respectively. Combined, they outcomes constitute 99% recognition rate on the selected type of actions. With the increase of the number of action or in case of changing the action type the accuracy of the system might decrease on 1- 1,5%.
4) The proposed system is able to capture, recognize and classify emotions and actions of a human subject in a robust manner. The integration of the emotion and action recognition allows to monitor human behavior in real time, providing more robust results than existing systems.
5) Experimental results demonstrate that the system is suitable for the implementation on the ATM machines. The system is ready for field tests and could be implemented for testing purposes in a typical ATM terminal.

Validation of the research results

The results were approved and published in the proceedings of the following **international conferences:**

1. The third conference of mathematical society of the Republic of Moldova. Chisinau: Institute of mathematics and Computer Science, Academy of Sciences of Moldova, 2014;
2. Development trends of contemporary science: visions of young researchers. Chișinau, Republic of Moldova, 2015;
3. Workshop on Foundations of Informatics - FOI-2015, August 24-29, 2015, Chisinau, Republic of Moldova;
4. The 7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016, March 8 - 11, 2016, Orlando, Florida, USA.

Publications on the thesis topic

Relying on the research results, **8 scientific papers** have been published (4 articles in reviewed scientific journals and 4 in conferences proceedings).

Thesis contents and structure. The thesis is written in English and typed at the computer as a manuscript. Thesis has the following structure: introduction, three chapters, general conclusions and recommendations, bibliography (109 sources). The thesis is presented in 121 pages of main text, 5 annexes, illustrated with 37 figures and 2 tables.

Keywords: computer vision; artificial neural networks; convolutional neural networks; emotion recognition; gesture recognition; radial basis function networks; self-organized maps; machine learning; infrared camera processing.

2. CONTENTS OF THE THESIS

In this work, we developed a neural network model for recognition of body movements and facial expression and for classification into two types (typical and non-typical). Such a complex task required both analysis of the emotional states of human subject, the whole spectrum of actions he performs in current situation and building and implementing of mathematical model, suitable for this task. We divided the overall problem into two sub-problems.

In the first part we have described the modular neural network, which we apply to the problem of emotion recognition. The architecture of basic module of the network is the self-organized map [5-7] (SOM) of functional radial-basis function (RBF) modules. Therefore, we provide a short mathematical introduction on this subject. In the first half of chapter two, we provide a detailed mathematical description of the applied approach. We formalize the mathematics of the model and provide an explanation on the choice and implementation of the model's learning algorithm. We demonstrate the implementation of the algorithm as a neural network model in chapter three.

The second part presents the description of convolutional neural network that we have used for the classification of action. We have used deep convolutional neural network for real-time classification of human body movements. We have resented the detailed mathematical notation of the architecture of the network, learning algorithm and the process of implementation and validation of the model.

The thesis **Introduction** describes the objectives of the research and outlines the main findings in the field of emotion and action recognition in terms of psychological issues and theoretical models. Here, we provide the reader with a brief overview of the system and it's major components. Also, we would outline such important issues, as relevance of the subject, the purpose and the objectives of the research and describe the methodology we use in the presented study. A description of the state of the art in the field of emotion and action recognition and identify of the research problems that exist in the field is provided too. The important scientific problem, which is solved in the research study, will be briefly described from the point of view of its theoretical significance and the applied value of the work. At the end of the introduction section, we provide the outline of the thesis with the detailed description of its three main chapters.

Such a general introduction is aimed to make the work accessible to a wide spectrum of readers, with different expertise and knowledge, since this work combines the results from both psychology and applied mathematics.

Chapter 1 (**Theory and models of emotion and action recognition**) is an introductory part. It contains description of the background and overview of the important studies on the related topics. In the first half of chapter 1, we provide the psychological basis for emotion and action recognition models. In the second half of chapter 1, we review the relevant literature on object recognition. During last decades, a large number of models of object recognition were proposed. They differ in many dimensions, for example in the number or type of emotions they recognize or in the machine learning techniques. In one manuscript, it is difficult to describe all the existing models, thus we will provide only the brief overview of the main types of object recognition models according to the approach they use. We will put a lot of emphasis on the existing mathematical approaches in neural network construction, which are close to the subject of this study.

Chapter 2 (**Neural network architecture and learning algorithms**)

Second chapter presents the architecture of the proposed neural network models for emotion and action recognition. This chapter could be divided into two logical parts. First part describes the modular neural network, which we apply to the recognition of emotions. The second part presents the description of convolutional neural network we use for the classification of gestures. First part describes the architecture of basic module of the network is the self-organized map (SOM) of functional radial-basis function (RBF) modules. We provide a mathematical introduction on this subject. In the first half of chapter two, we provide a detailed mathematical description of the applied approach: we formalize the mathematics of the model and provide an explanation on the choice and implementation of the model's learning algorithm.

We demonstrate the implementation of the algorithm as a neural network model in chapter three. In the second part of chapter two, we describe the second type of neural network (NN) architecture we use in our experiments. Second part describes an algorithm, which is able, with neither advanced pre-processing nor specific feature modelling or learning, to automatically extract and learn prominent features from the data as well as effectively classify them into one of the two gesture classes. In this part, we describe the architecture on deep convolutional neural network, which we use for classification: mathematical notation, learning algorithm and the process of implementation and validation of the model.

The architecture of our model is based on the notion of the self-organized map (SOM), proposed by Kohonen. This kind of neural network is trained using unsupervised learning to produce a two-dimensional map of the input space of the training samples. The quality of SOM to use a neighbourhood function for preserving the topological properties of the input space is used in our simulations to create the similarity map of the IT cortex. The conventional SOM algorithm has a number of restrictions, and the main one is its ability to deal only with the vectorised data. To solve this problem, a number of modifications of the conventional SOM have been proposed.

We used one of these modifications as a basis for constructing our model. This architecture has the number of advantages. First, every module in the modified SOM has the capability of information processing and can form a dynamic map that consists of an assembly of functional modules. Second, the RBF-SOM combines supervised and unsupervised learning algorithms: at the RBF-level, the network is trained by a supervised learning algorithm, i.e., the back propagation at the RBF module level, while the upper SOM level is described in an unsupervised manner. For the purposes of this study we used RBF network modules. The usage of RBFs instead of the MLPs adds the following properties to such a network while preserving the ability to form a dynamic map: 1) there is no need for an algorithm for avoiding local minima; 2) the network can recognize the object and store its representation in its inner centre. The generalized algorithm for processing the SOM of functional models can also be applied in this case. The architecture of the SOM of RBFs module has a hierarchical structure: it consists of two levels, which we will call the RBF-level and the SOM-level of the network. At the first level, the architecture of our network represents k RBF- networks, which are the modifications of the Poggio and Edelman network [8]. Since each module represents a certain "functional feature" determined by the model architecture, the SOM-level in the SOM of RBFs represents a map of those features. The proposed network solves an approximation problem in a high-dimensional space. Recognizing of an object is equivalent to finding a hyper-plane in this space that provides the best fitting to a set of training data. The training data represents a vector with coordinates of 2D projections of 3D objects, taken at each degree of rotation. In order to investigate the ability to classify complex 3D objects, such as faces, we extend our SOM of RBFs model by adding a hierarchical pre-processing module, presented in hierarchies of filters with the different degrees of resolution and pooling layers (Fig.1).

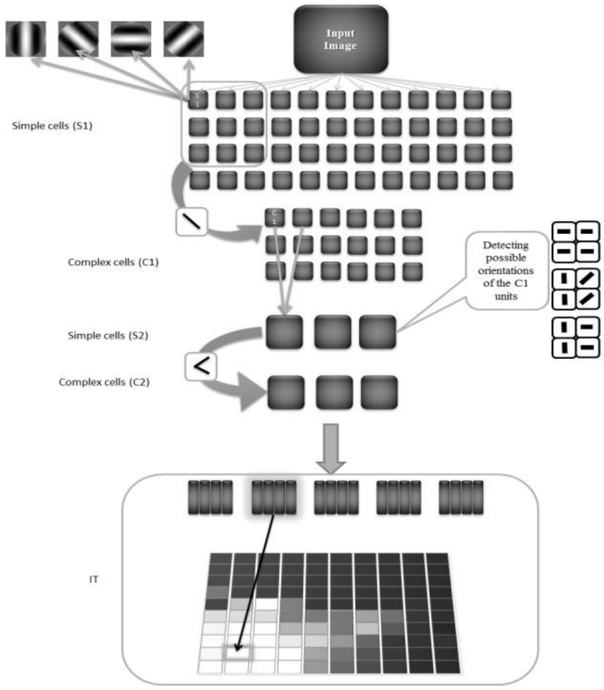

Figure 1. The architecture of the pre-processing module of simple and complex cells

The network output represents the activation map, the activation of each module shows the belonging of the detected expression to one of five basic emotions. for the purposes of this study, we selected five basic emotions, which are locates on a square plane, divided into 25 parts. The winning module represent the most plausible emotion. This approach allows defining the most plausible emotion or emotions (since the most active module can be defined between two emotions). In this study, we used only five emotions, but the usage of a larger number of emotion labels is also possible (Fig.2).

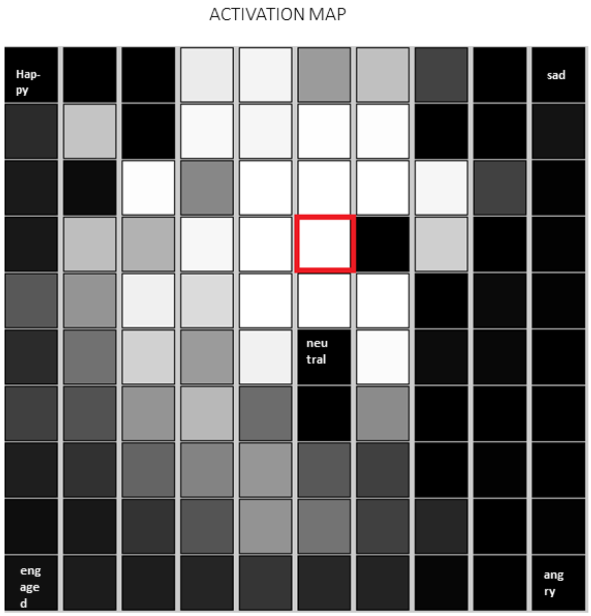

Figure 2. The output of RBFxSOM network

To address the problem of human actions recognition, we use convolutional neural network [9] (CNN) which architecture is extremely effective for classification of the large amount of data. By the term ‘actions’ here we understand the movements of the parts of the body that fell into the receptive field of the infrared camera, excluding face.

A deep neural network [10] (DNN) is an artificial neural network (NN) with multiple hidden layers of units between the input and output layers. Similar to shallow NNs, DNNs can model complex non-linear relationships. DNN architectures, e.g., for object detection and parsing generate compositional models where the object is expressed as a layered composition of image primitives. The extra layers enable composition of features from lower layers, giving the potential of modelling complex data with fewer units than a similarly performing shallow network.

The architecture of a CNN can be described as following. A small input region goes to input neurons and then connects to a first convolution hidden layer (Fig.3).

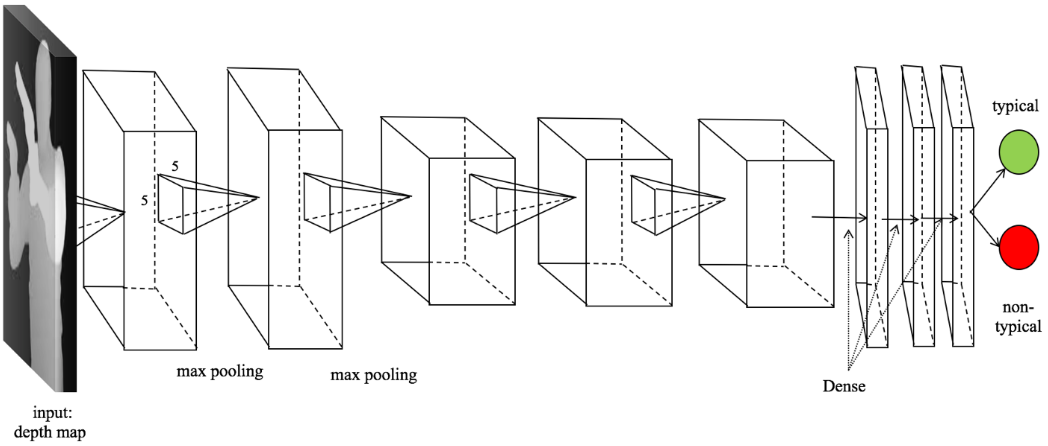

Figure 3. CNN architecture.

The input of the CNN is a normalized depth map, output is a classification of the input action (typical vs. non-typical). Between the input and output layer, we can see a set of learnable filters, which are activated during the presentation some particular type of feature in pixel region in the input. On this phase, CNN does shift invariance, which is carried by feature map. Subsampling layer goes next. There we have two processes: local averaging and sampling. As a result, we get declining resolution of feature map. In order to perform this task CNN needs supervised learning. Before starting the experiment, we gave a set of labelled videos with different emotional experience. The system analyses images and finds similar features. Then the system creates a map, where it arranges videos in accordance with similar features. Thereby, images with similar emotions form certain class. To test the system, we add other videos and correct the system when it refers them improperly.

The proposed model consists of three convolutional layers, followed by max-pooling layers, and three fully-connected layers with a final classificatory presented with MLP (with two basic outputs, corresponding to typical and non-typical behaviour). The input data was presented as filtered and normalized infrared camera output.

In order to process real time video, we propose a double-stream architecture, which incorporates spatial and temporal networks (Fig.4). Such a CNN, trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data, which is highly desirable in our case. This type of double-stream processing of video images was proposed in [11]. In the model, motion is explicitly represented using the optical flow displacement field, computed based on the assumptions of constancy of the intensity and smoothness of the flow.

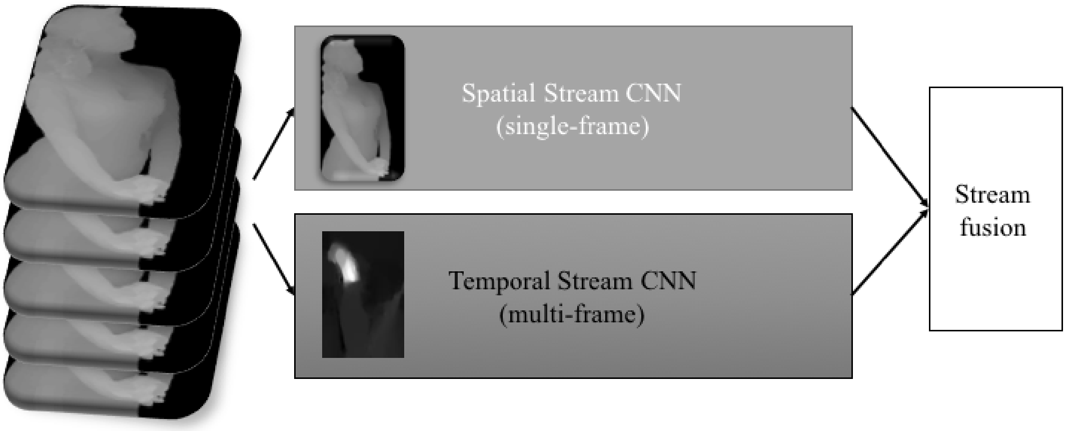

Figure 4. Double-stream architecture for video classification

For real time video processing, we suggest a double-stream architecture comprising spatial and temporal networks. It is obviously possible to separate spatial and temporal components in a video. The spatial part is responsible for carrying information about scenes and objects displayed in the video, which is reproduced through individual frame appearance. The temporal part reproduces the movement of the observer (the camera) and the objects by means of motion through the frames. Figure 4 shows that we develop our video recognition architecture according this principle, i.e. separating it into two streams. The use of softmax scores combined by late fusion in a deep CNN make each stream to be implemented. Averaging and training a multi-class linear SVM are regarded as the fusion methods.

The computations were performed on Python. The model was trained with the trained data and model evaluation was performed on the test data with the the k-fold cross-validation (for details, see next subsection). The computations were performed on the Amazon EC2 machine (https://portal.aws.amazon.com).

The reason for using external GPUs was the following. It was established, that using a single GTX 580 GPU with 3GB of memory only imposes strong limits on the maximum size of the networks that can be trained on it. For such large dimension of the trained network, as we have, it was necessary to use several GPUs. An important advantage of the nowadays GPUs lies in their ability of cross-parallelization, so that they can write to or read from one another's memory directly, avoiding addressing to the host machine. Following the well known experience we employ a specific parallelization scheme, which puts half of the neurons on each GPU, respecting the condition that the GPUs communicate only in certain layers. This means that, for example, the neurons of layer 3 take input from all kernel maps in layer 2. However, neurons in layer 4 take input only from those kernel maps in layer 3 which reside on the same GPU.

The communication model can be adjusted so that the connectivity amount became an acceptable value from the total volume of computation.

The validation of the neural network model was performed with the leave one out cross validation (LOOCV) technique. The use of LOOCV was essential for appropriate estimation of optimal level of regularization and parameters (connection weights) of neural network obtained. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. LOOCV is a particular case of leave-p-out cross-validation. Leave-p-out cross-validation (LpOCV) involves using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of p observations and a training set. LpO cross-validation requires to learn and validate *Cnp* times (where n is the number of observations in the original sample). In Leave-one-out cross-validation we assume $p = 1$. However, for our purpose LOOCV appeared to be relatively slow.

Therefore, the validation of the CNN network results was performed with the K-fold cross-validation technique [12]. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter. When $k=n$ (the number of observations), the k-fold cross-validation is exactly the leave-one-out cross-validation. The results of two NN outputs were fused together in a rule-based manner and allowed us to combine the outputs of two separate sub-systems.

Chapter 3 (**Research applications and psychological experiments**) contains two consistent sections: description of experimental sessions for emotion recognition and for action recognition, respectively.

Most research is focused on detecting facial expressions in an isolated framework, where each target is analyzed separately. Here we present a collective framework to analyze the group emotions and general human behavior.

The aim of our research is to use the infrared cameras for capturing the image of the user. Here, we use Kinect API to record a database of multiple targets. However, other types of cameras could be used for these purposes. A ground truth database containing manually labelled emotions will be also created for analysis and evaluation purposes.

We propose a hybrid architecture for complex event analysis. The real-time analysis of human reactions (facial expression and gestures) is performed with the help of state-of-the art machine learning techniques, described in chapter two. The resulting measurements are compared with the statistical data, recorded earlier and human observer's data.

For the purpose of the study, we have mounted the hardware stand, which consisted of the ATM terminal with mounted infrared camera on the top of this terminal. We have written the simulator of the card processing, similar to the one which is usually used in the ATM machines. The software allowed the user to perform one of the four standard operations: money deposit, money withdrawal, money transfer to other account and putting the money on the cell phone account. The data for the software was taken from a few popular banks and averaged to represent the "standard ATM machine". The software was written on Java.

The users were divided into two groups: positive test examples and negative test examples. The subjects from the "positive" group had to perform standard actions with the ATM in the way it they usually do, without any additional recommendations. Using this hardware-software complex, the human subject, who participated in the experiments, was asked to perform a number of actions (money deposit, money withdrawal, money transfer to other account and putting the money on the cellphone account). In the "negative" example group users were asked to perform non-standard actions (trying to hack the machine, pretending to be drunk, pretending to rob another user etc.).

All the actions were recorded on video and presented to the group of observers, who labelled the actions as being "typical" or "non typical". The obtained data was used further for training the neural network.

The group of test users were asked to perform the standard actions with the ATM to verify the performance of the system in real time.

The visual information in the proposed system is presented in several steps:

1. First, we use cameras and 3D sensors such as the Microsoft Kinect to detect facial features in order to recognize and classify emotions and gestures.
2. Second, we apply computer vision techniques for feature extraction and pattern recognition.
3. We apply machine learning (neural networks) for emotion detection and classification.
4. We use recorded statistical data from the machine transactions or logs for the training of our system. We train a modular neural network together with the emotion records to provide the analysis of the events. We can use the trained networks for real-time analysis of user's actions.
5. During the interaction of the user with the system we can track fraudulent actions in real-time and initialize security measures in order to prevent crime or fraud.

Two series of experiments were conducted. First group of experiments was conducted in order to evaluate how effectively the proposed system can detect normal vs. abnormal behavior of customer during interaction with ATM. For the purposes of experiment, we developed an ATM simulation software, that was used in the stand-alone terminal. During the interaction session, the reactions of users were recorded by a camera, mounted on the top of the terminal.

The obtained records were later evaluated by human observers; and emotions, displayed on these records, were classified as typical or non-typical. In order to record the emotions, which were not displayed (according to human observers and subjective feelings of the participants of the experiment) during the first series of experiments, we recorded the emotions, displayed by the same subjects during the observation of short videos. In order to preserve the uniformity of the data, we showed the videos on the same equipment which were used during the ATM-experiment session.

Twenty healthy subjects, age 21-37, with normal or corrected-to-normal vision, participated in the experiment. Simultaneously, the data from two series of experiments was processed with an infrared camera and used as an input to the neural network model. Each subject performed 10 sessions with the ATM-simulation software and five video session.

During the interaction session, the reactions of users were recorded by a camera, mounted on the top of the terminal. The receptive field od the camera includes whole body, from the head on the top till the knees on the bottom. However, for the purposes of this part of experiments, only the face was processed and analyzed (upper right corner of the figure).

In order to evaluate the performance of the neural network model, we run the simulation experiments with the same input data from infrared cameras. Also, we used the data from the same human subjects, displaying other emotions. Total 7 emotions were displayed by each subject.

The resulting recordings were randomly classified into training and testing subsets. During the simulations, the network classified the "typical" behavior of the ATM used with the 86% accuracy.

In the second series of experiment, we used one of the approaches to recognition of gestures is body tracking: classification of the body movements. One of the the classification techniques for this method is pattern recognition: i.e. special video/infrared camera recognizes human actions: waving, jumping, hand gestures etc. Among the first successful representatives of this technology are Kinect from Microsoft. The Kinect uses structured light and machine learning as follows:

- The depth map is constructed by analyzing a speckle pattern of infrared laser light.
- Body parts are inferred using a randomized decision forest, learned from over 1 million training examples.
- Starts with 100,000 depth images with known skeletons (from a motion capture system).
- Transforms depth image to body part image.
- Transforms the body part image into a skeleton.

For the purposes of the study, we do not use the classification technique, proposed by Kinect, but use it only as an infrared sensor.

Psychological experiments. We conducted a series of experiments in order to evaluate how effectively the proposed system can detect normal vs. abnormal behavior of customer during interaction with ATM. For the purposes of experiment, we developed an ATM simulation software that was used in the stand-alone terminal. During the interaction session, body movements of users and facial expressions were recorded by a camera, mounted on the top of the terminal. These records were later evaluated by human observers; and behavior, displayed on these records, were classified as typical or non-typical. In order to preserve the uniformity of the data, we showed the videos on the same equipment which were used during the ATM experiment session. Thirty healthy subjects, age 21-37, with normal or corrected-to-normal vision, participated in the experiment. Simultaneously, the data from two series of experiments was processed with an infrared camera and used as an input to the CNN algorithm. Each subject performed 10 sessions with the ATM-simulation software and 5 video session. During each session, the recognition of the upper-body movements (in the range of the camera, mounted on the top of the typical ATM machine) was performed together with facial features classification and recognition. Among thirty subjects, we used 22 as examples of 'normal' behavior and 8 as examples of 'abnormal' behavior. Figure 5 demonstrates the data samples, obtained during our experiments.

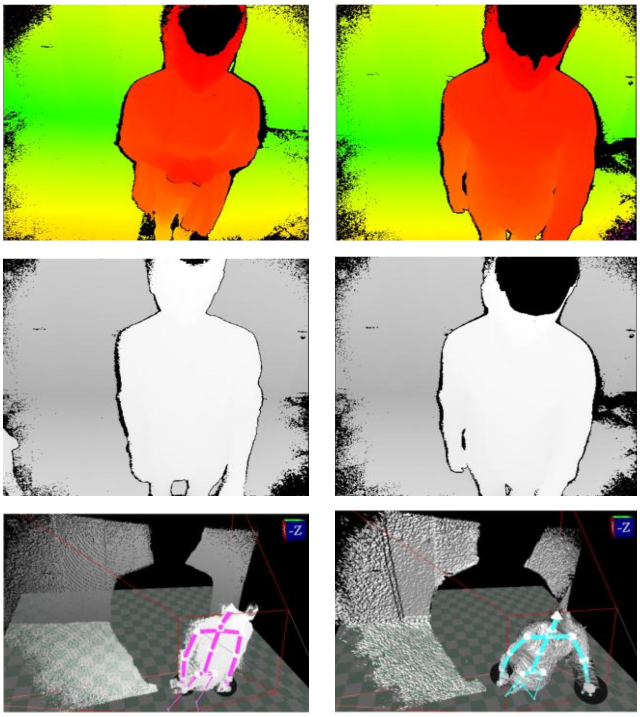

Figure 5. Data samples, obtained during our experiments.

Screenshots are captured from the infrared camera, mounted on the top of the ATM terminal and showing the human subject from the position, in which the typical surveillance camera would be mounted. Screenshots are captured from the infrared camera, mounted on the top of the ATM terminal, showing two consecutive actions: hands down (no action) and one hand up (entering the pin-code).

The overall system performance is described in terms of the output: whether it classifies the user's behavior (emotions + gestures) as typical or non-typical.

We need to emphasize, that several combinations of the algorithms as possible. RBFxSOM is, generally speaking, a simpler algorithm than CNN. We use it parallel with CNN only for two main purposes:

- By employing this algorithm, we obtain a continuous map of features, which is easier to interpret in comparison to classification on only two classes of emotions (typical – Fig.6 vs. non-typical – Fig.7).
- We don't have to train the CNN twice, which computationally is cheaper.
- However, this step could be omitted if we have enough time and capacity to train CNN twice: for both emotions and body movements classification.

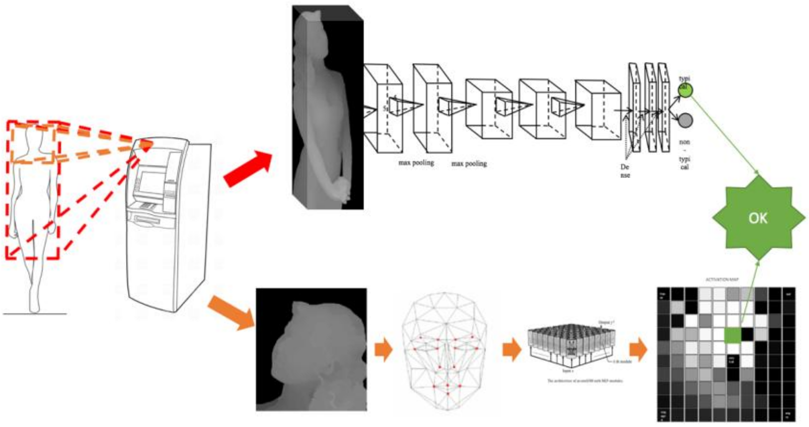

Figure 6. The real-time system's feedback on user's behavior. Case A: the behavior of the user is classified as "typical".

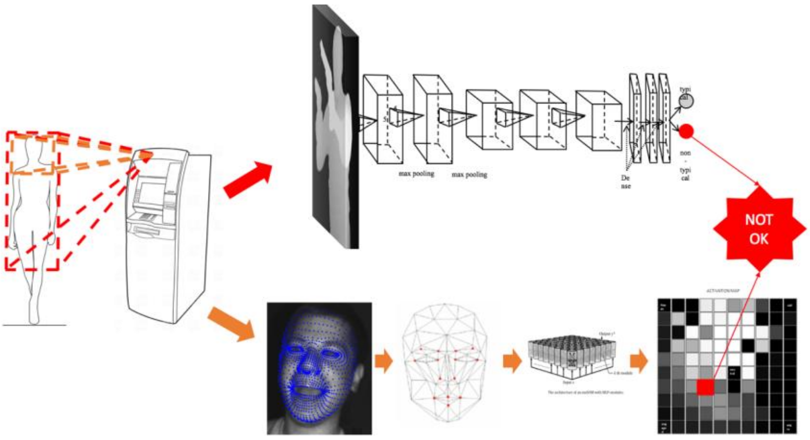

Figure 7. The real-time system's feedback on user's behavior. Case B: the behavior of the user is classified as "non-typical".

Therefore, we have build and tested the complex video security system, suitable for the ATM machines. The system is ready for field tests and could be implemented for testing purposes in a standard ATM terminal.

3. CONCLUSIONS AND CONTINUATIONS

1) We have proposed a tool for robust classification of emotions and gestures of a human subject into typical vs. non-typical in a certain kind of interaction. We described two types of neural network architectures for classification of human gestures and emotions, obtained from infrared cameras. These architectures could be used in parallel (for more fast and robust processing), or only CNN for feature processing can be used. The choice of architecture depends on the circumstances, in which the system can be used. This study can be regarded as an attempt to make one more step towards the implementation of this kind of architectures: we apply two types of hierarchical modular architectures to the task of recognition of human emotions and actions and use it to solve the real problem of classification of human behaviour into proper and improper for a certain task.

2) The proposed approach can be used in a variety of applications. We restrict ourselves to only one type of interaction (user of an ATM machine) for simplicity. However, this kind of classification task is very useful in the number of applications, where the number of gestures of the human is limited, such as customers at the various types of automated machines. For this category of users, the algorithm can be used for detection of unusual/fraudulent behaviour to decrease the workload of the closed-circuit television (CCTV), or video surveillance, operators who monitor users of these machines. For example, this system could be applied for monitoring workers in surrounding, where their actions are significantly restricted: assembly line, construction works on high buildings, in the underground, in mines). Another example will be monitoring driver’s/pilot’s arousal, attention etc. With the help of this system, we could classify correct vs. incorrect actions, identify such unwanted stated as loss of attention, sickness, tiredness etc.

3) The results, presented in this work, demonstrate that the proposed system maintain the recognition rate similar to the state of the art in the computer vision and emotion recognition fields. Moreover, the system of such complexity, incorporating both emotion and action recognition, has not been presented so far. The architecture performs recognition of body movements and facial expression and for classification into two types (typical and non-typical) with the overall accuracy (of 8% and 14% error rate, respectively). These results were achieved

independently and combined afterwards with a simple classification rule-based algorithm. The combination of the results improves the subsystem's performance if they work independently.

4) To improve system's performance, the proposed model requires a large amount of training data, which cannot be easily obtained. Therefore, the natural continuation of current research would be conducting further field tests to obtain more training data and improve performance.

5) Experimental results demonstrate that the system is suitable for thee implementation on the ATM machines. The system is ready for field tests and could be implemented for testing purposes in a typical ATM terminal.

Therefore, we can conclude that all the goals of the current research were obtained and the technical sub-tasks were successfully implemented. We can conclude that the proposed system is able to:

- capture, recognize and classify emotions and actions of a human subject in a robust manner;
- the integration of the emotion and action recognition allows to monitor human behaviour in real time, providing more robust results than existing systems.

Future work

The research, described in this study, constitutes the tiny part of the spacious area of human recognition. We have only touched the possibilities, which open at the moment by the implementation of deep learning and neural network systems in the industrial applications. The natural continuation of this research will be the construction of the neural network with a wider range of actions it would be able to recognize. Such a network would be suitable for applications, where the number of gestures of the human is limited, such as customers at the various types of automated machines, CCTV or video surveillance systems, operators who monitor users of ticket machines at the underground station and self-checkout cashiers, long-distance drivers, assembly line workers, construction workers and many others.

REFERENCES:

1. Ekman P. Basic Emotions. In: Handbook of Cognition and Emotion. New York, NY: John Wiley and Sons Ltd., 1999, ch. 3, p. 45–60.
2. Ekman P., Friesen W. Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, 1978.
3. Albu V., Cojocaru S. Measuring human emotions with modular neural networks and computer vision based applications. In: Computer Science Journal of Moldova, 2015, vol.23, vol. 1, no. 67, p. 40-61.
4. Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Fei-Fei L. Large-scale Video Classification with Convolutional Neural Networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, p. 1725-1732.
5. Kohonen T. Self-organizing maps. Series: Springer Series in Information Sciences, vol. 30, Berlin: Springer-Verlag, 2001. XX+502 p.
6. Furukawa T. SOM of SOMs. In: Neural Networks, May 2009, vol. 22, no. 4, p. 463-478.
7. Tokunaga K., Furukawa T. Modular network SOM, In: Neural Networks, January 2009, vol. 22, no. 1, p. 82-90.
8. Poggio T., Edelman S. A network that learns to recognize three-dimensional objects, In: Nature, 1990, vol. 343, p. 263 - 266.
9. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation Applied to Handwritten Zip Code Recognition, In: Neural Computation, 1989, vol. 1, p. 541-551.
10. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, In: Biological Cybernetics, 1980, vol. 36, no. 4, p. 193-202.
11. Albu V. Measuring human emotions with modular neural networks. In: The proceedings of the 7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016, March 8 - 11, 2016, Orlando, Florida, USA, p. 26-27.
12. Kamnitsas K., Chen L., Ledig C., Rueckert D., Glocker B. Multi-Scale 3D Convolutional Neural Networks for Lesion Segmentation in Brain MRI. In: Proceedings of MICCAI Brain Lesion Workshop 2015, Munich, Germany, 2015. http://hdl.handle.net/10044/1/27804 (visited on March, 29, 2016)

AUTHOR’S PUBLICATIONS ON THESIS SUBJECT

1. Albu,V. Measuring customer behavior with deep convolutional neural networks. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, Volume 1, Issue 2 , April 2016, pp.74-79. E-ISSN 2067 – 3957 **(ISI).**
2. Албу, В.А., Хорошевский, В.Ф. КОГР система когнитивной графики. Разработка, реализация и применение. В: Известия Академии Наук СССР. Техническая кибернетика. 1990, nr. 5, pp. 105-118.
3. Averkin, A., Albu, V., Ulyanov, S. and others. Dynamic object identification with SOM-based neural networks. In: Computer Science Journal of Moldova, 2014, nr. 22 1/64, pp. 110-126. ISSN 1561-4042 **(B+)**
4. Albu, V.; Cojocaru, S. Measuring human emotions with modular neural networks and computer vision based applications, *Computer Science Journal of Moldova,* v.23, n.1 (67), 2015, pp.40-61. ISSN 1561-4042 **(B+)**
5. Albu,V. Neural network based model for emotion Recognition. In: *Proceedings of the Workshop on Foundations of Informatics.* FOI-2015, August 24-29, 2015, Chisinau, Republic of Moldova, pp.423-434
6. Ulyanov, S., Albu, V., Barchatova, I. Intelligent robust control system based on quantum KB-self-organization: quantum soft computing and Kansei / affective engineering technologies. The third conference of mathematical society of the Republic of Moldova. Chisinau: Institute of mathematics and Computer Science, Academy of sciences of Moldova, 2014, pp. 571-582. ISBN: 978-9975-68-244-2.
7. Albu V. Measuring human emotions with modular neural networks. In: Proceedings of the 7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016, March 8 - 11, 2016, Orlando, Florida, USA, pp.26-27.
8. Albu, V. Measuring human emotions with modular NNS and computer vision applications. În: *Tendințe contemporane ale dezvoltării științei: viziuni ale tinerilor cercetători.* Teze ale Conferinței Științifice Internaționale a Doctoranzilor. Martie, 2015, AȘM, Chișinău, p.14.

ABSTRACT

of the thesis "Human actions recognition with modular neural networks" submitted by Veaceslav Albu for fulfillment of the requirements for the Ph.D. in Computer Science, specialty 122.03 – Modeling, mathematical methods, software. The thesis was elaborated at the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova, Chisinau, in 2016. The thesis is written in English and contains Introduction, 3 chapters, general conclusions and recommendations, bibliography of 109 titles. The main text amounts to 121 pages. This work includes: 37 figures, 2 tables, and 5 annexes. The results are published in 8 scientific papers.

Keywords: Deep Neural Networks, Computer Vision, Emotion Classification, Gesture Classification.

The area of the present studies is the field of emotion and action recognition using modular neural networks.

The aim and objectives of this research is to develop a tool for classification of human reactions (including facial features and body movements) into typical and non-typical in a certain environment. This tool provides statistical observations and measurements of human emotional states during an interaction session with a software product (or, optionally, with a hardware plus software complex).

Scientific novelty is a novel modular neural network architecture, constituted from two separate parts and combine the results to introduce the classification of the infrared sensor inputs, which is the first system of this kind, being applied both to emotion and human action recognition.

The important solved scientific problem is elaboration of a multimodal method for classification of human reactions (joining emotions and actions) into typical and non-typical in a certain environment, that ensures an effective functioning of systems destined to human actions monitoring in real time.

Theoretical significance. Our research solutions provide ground for solving of following problems: formulation of the tool's architecture for robust classification of emotions and gestures of a human subject into typical vs. non-typical; the substantiation of the possibility and efficiency of using deep learning in an integrated approach for the detection of expression of the whole body in real time**.**

Practical value: this kind of classification task is very useful in different applications, where the number of gestures of the human is limited, such as: customers at the various types of automatedmachines, drivers, assembly line workers, hospital patients etc.

ADNOTAREA

tezei “Recunoașterea acțunilor umane în baza rețelelor neurale modulare” înaintate de către Albu Veaceslav pentru obținerea titlului de doctor în informatică la specialitatea 122.03 – Modelare, metode matematice, produse program. Teza a fost elaborată în Institutul de Matematică și Informatică al AȘM, Chișinău, anul 2016. Teza este scrisă în limba engleză și constă din introducere, trei capitole, concluzii generale și recomandări, bibliografie ce cuprinde 109 titluri. Lucrarea conține 121 pagini text de bază, 37 figuri, 2 tabele, 5 anexe. Rezultatele principale sunt publicate în 8 lucrări științifice.

Cuvinte cheie: Rețele neurale adânci, computer vision, clasificarea emoțiilor, clasificarea gesturilor.

Domeniul de studiu al tezei îl constituie rețelele neurale modulare.

Scopul și obiectivele cercetării ține de elaborarea unui instrumentar pentru clasificarea reacțiilor umane (care includ aspecte faciale și mișcări ale corpului) în două clase: tipice și atipice pentru anumit mediu. Acest instrument oferă observații și măsurări statistice ale stărilor emoționale umane în timpul unei sesiuni de interacțiune cu un produs software (sau, opțional, a interacțiunii cu un complex hardware și software).

Noutatea și originalitatea cercetării o constituie arhitectura nouă a rețelei modulare neurale, care constă din două părți separate, combinându-le rezultatele în scopul efectuării unei clasificări a datelor obținute de la sensori infraroșii. Acesta este un prim sistem de acest fel aplicat atât pentru recunoșterea emoțiilor faciele, cât și a acțiunilor umane.

Problema științifică importantă soluționată constă în elaborarea unei metode multimodale de clasificare a reacțiilor umane (unind emoțiile și acțiunile) în tipice și atipice în raport cu un mediu dat, fapt care asigură funcționarea eficientă în timp real a unor sisteme de monitorizare a acțiunilor umane.

Semnificația teoretică. Rezultatele cercetării fundamentează soluționarea următoarelor probleme: stabilirea arhitecturii instrumentarului pentru clasificarea fiabilă a emoțiilor și gesturilor a unui subiect uman în tipice vs. atipice; stabilirea posibilității și eficienței utilizării învățării profunde în cadrul unei abordări integrate pentru identificarea expresiilor întregului corp uman în timp real.

Valoarea practică: soluționarea acestei probleme de clasificare este extrem de utilă pentru diverse aplicații, în care numărul de gesturi umane este limitat, precum cel al utilizatorilor mașinilor automate de cel mai variat tip, conducători auto sau cei de trenuri, muncitori la linii de asamblare, pacienți în spitale, aflați în stare de imobilitate etc.

АННОТАЦИЯ

диссертации “Распознавание действий человека на основе модулярных нейронных сетей” представленной Вячеславом Албу на соискание ученой степени доктора наук в области информатики по специальности 122.03 – Математическое моделирование, методы, программное обеспечение. Диссертация была написана в Институте математики и информатики при Академии наук Молдовы (Кишинёв), в 2016 году, на английском языке и содержит: введение, три главы, общие заключения и рекомендации, библиографию из 109 названий, 121 страницу основного текста, 37 рисунков, 2 таблицы, 5 приложений. Полученные результаты опубликованы в 8 научных статьях.

Ключевые слова: глубинные нейронные сети, компьютерное зрение, классификация эмоций, классификация жестов.

Областью исследований диссертации являются модулярные нейронные сети.

Целью диссертации является разработка инструментария для классификации реакций человека (включающих в себя выражение лица и движения тела) на два вида: типичные и нетипичные для определенной среды. Этот инструментарий предоставляет возможность проведения статистических наблюдений и измерений эмоционального состояния человека при его взаимодействии с некоторым программным комплексом (или, как вариант, с аппаратно-программным комплексом).

Научная новизна и оригинальность диссертации выражены в новой архитектуре модулярной нейронной сети, которая состоит из двух отдельных частей, результаты которых объединяются для осуществления классификации данных, полученных от инфракрасных датчиков. Это первая система такого рода применяемая как для распознавания лицевых эмоций, так и человеческих действий.

Решена важная научная проблема, которая заключается в создании мультимодального метода классификации человеческих реакций (объединяющих эмоции и действия) на типичные и нетипичные по отношению к данной среде, что обеспечивает эффективное функционирование в режиме реального времени систем мониторинга человеческих действий.

Теоретическая значимость полученных результатов состоит в обосновании решения следующих задач: создание архитектуры комплекса для надежной классификации действий на типичные и нетипичные, доказательство возможности использования глубинного обучения в рамках интегрированного подхода для распознавания выражений человеческого тела в целом в режиме реального времени.

Прикладная ценность: решение задачи классификации находит применение в ряде приложений, в которых количество жестов ограничено, например различного типа автоматы, водители автомобилей или поездов, работники сборочных линий, пациенты, находящиеся в неподвижном состоянии.

ALBU VEACESLAV

HUMAN ACTIONS RECOGNITION WITH MODULAR NEURAL NETWORKS

SPECIALTY: 122.03

MODELING, MATHEMATICAL METHODS, SOFTWARE

Abstract of the Ph. D. Thesis in Computer Science

Aprobat spre tipar: 30.05.2016	Formatul hârtiei 60x84 1/16
Hârtie ofset. Tipar ofset.	Tiraj 50 ex.
Coli de tipar 1.7	Comanda nr.

Valinex S.R.L.
Str.Florilor, 30/A, of.26b, mun. Chişinău, MD 2068

ACADEMIA DE ȘTIINȚE A MOLDOVEI
INSTITUTUL DE MATEMATICĂ ȘI INFORMATICĂ

Cu titlu de manuscris
C.Z.U: 519. 95

ALBU VEACESLAV

RECUNOAȘTEREA ACȚIUNILOR UMANE ÎN BAZA REȚELELOR NEURALE MODULARE

SPECIALITATEA: 122.03
MODELARE, METODE MATEMATICE, PRODUSE PROGRAM

Autoreferatul tezei de doctor în informatică

Chișinău, 2016