Semantic analysis of the video stream based on neuro-symbolic artificial intelligence

Kroshchanka Aliaksandr, Mikhno Egor
Brest State Technical University
Brest, Belarus
kroschenko@gmail.com,
dzinushi.kun@gmail.com

Kovalev Mikhail, Zahariev Vadim,
Zagorskij Alexandr
Belarusian State University
of Informatics and Radioelectronics
Minsk, Belarus
michail.kovalev7@gmail.com, zahariev@bsuir.by,
alexandr.zagorskiy.work@gmail.com

***Abstract*—In the article, the model developed by the authors is considered, which is used for the semantic analysis of the video stream. The model is based on a neuro-symbolic approach. The features and advantages of the model are described. Based on the proposed model, a hybrid system for semantic analysis of the emotional state of the user is implemented. The configuration of the hardware platform necessary for the operation of the developed system is given.**

***Keywords*—neuro-symbolic AI, computer vision, artificial neural network, knowledge base, inference**

I. Introduction

In the last decade, there has been a steady tendency to widely use methods of machine learning and computer vision in various areas of human activities, primarily due to the development of the theory of artificial neural networks and hardware capabilities.

The development of applied methods in the field of computer vision leads to new original practical solutions.

The number of processes, which are being automated using new approaches in computer vision and which often could not be automated with acceptable quality earlier, is growing rapidly. In industries, it has become possible to reduce human participation in the process of product development and quality control [1]; in medicine, computer vision is used to analyze medical images; in the transport industry, it helps to carry out visual control of maintaining traffic regulations and operate autonomous vehicles.

Developments that are able not only to support basic functionality (even intellectual one) but also to conduct complex semantic analysis that produces new knowledge that can be used to improve the quality of the system as a whole are of incontestable value.

This combination involves the joint usage of ideas and methods from the fields of artificial neural and semantic models (and in the limit – connectionist and symbolic approaches in AI).

The advantage of artificial neural networks (ANN) is that they can work with unstructured data. The main disadvantage of ANN is the lack of human-understandable feedback, which could be called a reasoning chain, i.e., it can be said that ANN work as a “black box” [2].

Symbolic AI is based on symbolic (logical) reasoning. Such AI allows solving problems that can be formalized and plays an important role in human knowledge [3], [4]. However, it is not designed to work with unstructured data. Thus, a proper combination of these approaches will allow transforming unstructured data into knowledge.

The building of hybrid neuro-symbolic systems has already been widely covered in the literature [5]–[8], but currently all these attempts are mostly theoretical.

In this article, a neuro-symbolic model is proposed, which integrates various ANN models, engaged in solving problems of detection and recognition, with a knowledge base built on the basis of an ontological approach [9]. Based on the proposed model, the authors developed a real system for semantic analysis of the emotional state of the user.

The article is structured as follows:

- in section II, the problem definition to develop a model of a computer vision system for the semantic analysis of a video stream based on a neuro-symbolic approach is presented;
- section III is dedicated to the architecture of the developed model and the integration of its main components;
- section IV is concerned with the description of aspects of the implementation of a hybrid system for the semantic analysis of the emotional state of the user based on the proposed model;
- finally, in section V, the description of the hardware platform, on which the system described in section IV is launched, is given.

II. Problem definition

The authors were set the problem of developing:

1) a model of a computer vision system designed for semantic analysis of a video stream;

2) a hybrid system for semantic analysis of the emotional state based on this model.

When defining the problem, it was assumed that the nature of the data analyzed by the model can change, so the model should have a modular structure that allows for the simple replacement of modules in case of changes in the analyzed data.

The model should analyze the most semantically “rich” high-level data that can be obtained as a result of the detection of any objects and their recognition.

The following requirements are imposed on the model:

- support for various embedded modules for detecting and recognizing objects on video;
- simplicity in adding new models;
- semantic analysis of the results of recognition;
- availability of using the results of semantic analysis;
- the possibility of explaining the results of this analysis to a human.

To meet the specified requirements, a model of neuro-symbolic AI is proposed, within which the interaction of ANN and KB is organized to solve the problems of image recognition and semantic analysis, respectively.

The following requirements are imposed on the project of a hybrid system for semantic analysis of the emotional state of the user:

- The system should be able to identify the face of a person from the list of persons known to it;
- The system should be able to determine the fact of the appearance of an unknown person and add it to the list of persons known to it;
- The system should evaluate the emotional state of the person in the frame and respond accordingly;
- The system should accumulate statistical information about persons, who appeared in front of the camera, and their emotional state.

All the specified problems were solved within the developed model and the system based on it.

III. Overall structure of the proposed model

The main components of the proposed system model are:

- an interface, which in the simplest case can be represented by a camera, the video stream from which is transmitted to the computer vision module;
- a computer vision module, in which the input video stream is split into separate frames, which are recognized by available neural network models, and the obtained recognition results are transmitted to the knowledge integrator;
- a knowledge integrator that forms the necessary constructions from the results of recognition and places the generated knowledge in the knowledge base;
- a knowledge base that stores knowledge about the subject domains of the problems of recognition solved by the system, indications of neural network models that solve these problems as well as the results of recognition;
- a problem solver that performs semantic analysis of the results of recognition in the knowledge base.

Figure 1 shows the scheme of interaction of the above components. This scheme displays the upper layer of abstraction of the system components, which in practice can be implemented more comprehensively. The interaction between the above components can be implemented in various ways, for example, based on a network protocol, direct access within a monolithic architecture, shared access to some data storage, etc.

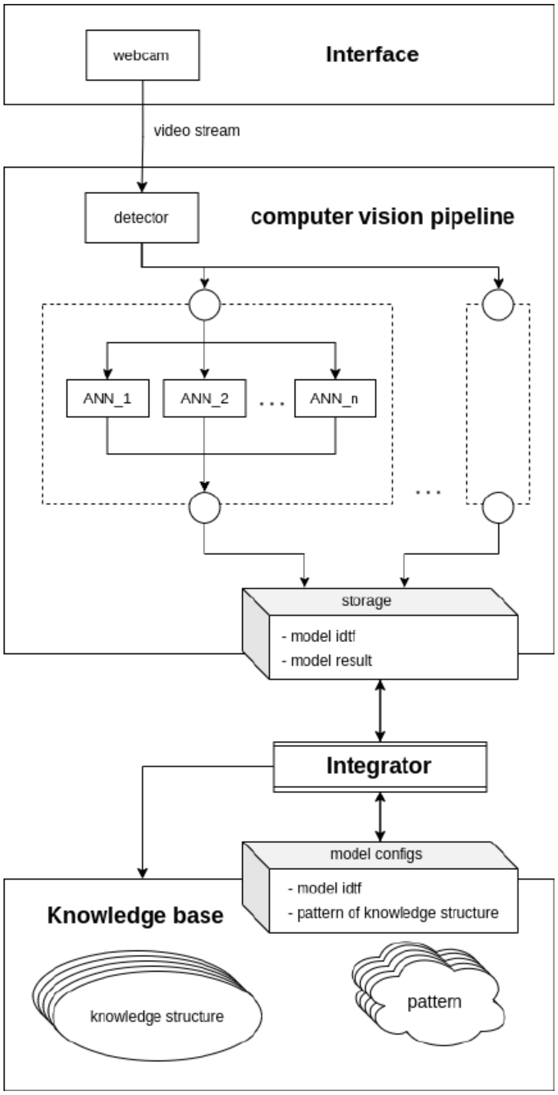

Figure 1. The scheme of interaction of the main components of the system

Since the developed model is a hybrid, the mechanisms of integration of the computer vision module and the knowledge base are of the greatest interest. Next, let us consider these mechanisms in more detail.

A. Integration of the computer vision module with the knowledge base

The integration of the computer vision module with the knowledge base as well as the direct implementation of the knowledge base and its processing are based on the usage of the OSTIS Technology [10].The model of the neuro-symbolic AI system considered in this article is an ostis-system.

Among the advantages of ostis-systems, it is possible to distinguish:

- the ability to perform semantic integration of knowledge in own memory;
- the ability to integrate different types of knowledge;
- the ability to integrate various problem-solving models.

The internal language used by ostis-systems is called an SC-code. Any knowledge can be represented in the form of an SC-code construction. A set of knowledge forms the ostis-system knowledge base. It is quite convenient to operate with knowledge in such a form: the technology supports the search and generation of necessary constructions both according to special templates and elementwise. You can read more about knowledge representation using the SC-code here [11].

1) Knowledge integrator: To carry out a semantic analysis of the result of the functioning of the neural network model, it is necessary to place these results in the knowledge base, i.e., to transform information into knowledge. The problems of converting information into knowledge are:

- the design of knowledge structures (defining the key nodes of the selected subject domain, their coordination with the formed knowledge base);
- automation of the formation and placement of structures in the KB.

Within the framework of the computer vision module, these problems are solved by the knowledge integrator, which places the results of the functioning of neural network models in the KB.

Neural network models solve the following problems:

- detection of objects;
- identification of detected objects of a certain class;
- recognition of the class of all detected objects.

As a result of solving these problems, various knowledge is placed in the knowledge base:

- knowledge about the presence/absence of objects in the "field of vision";
- knowledge about the correspondence of the detected objects of a certain class to some entities available in the knowledge base (if there are no such entities, new ones are created);
- knowledge about the classes of detected objects and the periods, during which the objects corresponded to these classes.

Due to this unification of the problems that neural network models solve within the computer vision module, it is possible to achieve the independence of the operation of the integrator from particular neural network models. It would be adequate to formalize the specification of such models in the knowledge base, which includes:

- the type of problem being solved;
- input and output data (depending on the type of the problem);
- the work frequency (if necessary);
- the state (on/off);
- the identifier, by which the specification of the model in the knowledge base can be correlated with the model in the computer vision module.

The knowledge integrator receives the specifications of all neural network models presented in the computer vision module and performs the integration of knowledge in accordance with these specifications.

Figure 2 shows an example of formalization of the specification of a neural network model for recognizing emotions in the knowledge base.

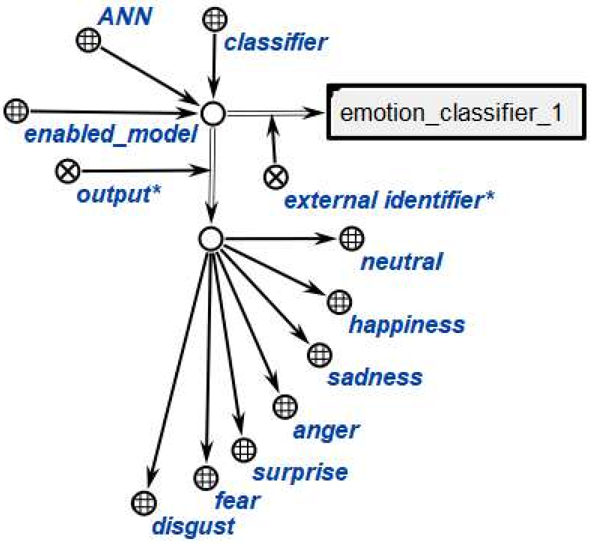

Figure 2. An example of formalization of the specification of a neural network model for recognizing emotions in the knowledge base

This neural network model solves the problem of classifying emotions and outputs for each recognized class a degree of confidence that the recognized object belongs to this class. Further, in accordance with the period stated in the model specification, statistics of the operation of such a neural network model are accumulated.

Depending on the configuration of the neural network model, which indicates the need to place operation statistics, confidence percentage for the answers, active time, etc. into the knowledge base, the integrator chooses a template, according to which it will generate knowledge. Such templates are also presented in the knowledge base.

Figure 3 shows an example of a template for generating the result of recognizing an object class over a certain period.

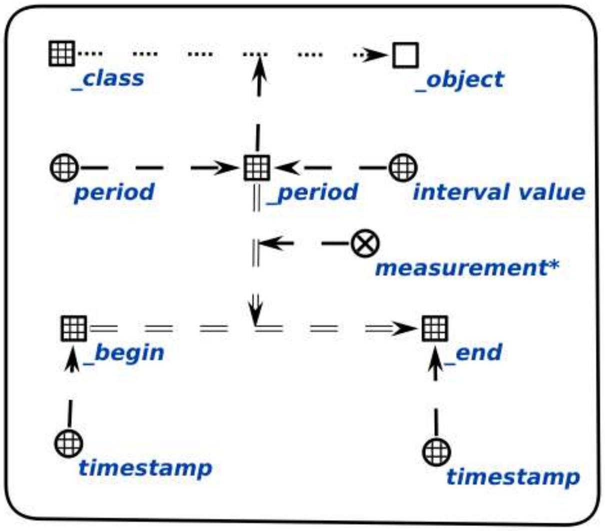

Figure 3. A template for generating the result of recognizing an object class over a certain period

An example of the knowledge structure that is formed based on the results obtained by the neural network model of emotion recognition of the computer vision module is shown in figure 4.

The approach, when the integrator works with the specifications of neural network models in the knowledge base, allows:

- avoiding overhead costs when integrating new neural network models, since it is enough only to describe the specification of the new model in the knowledge base;
- managing the computer vision module from the KB (for example, it is sufficient to add a neural network model to a set of enabled models so that the computer vision module stops running this neural network model for a video stream).

IV. Implementation of a hybrid system for analyzing the emotional state

Let us consider in more detail the architecture of the computer vision module and the principle of operation of the semantic analyzer.

A. Structure of the computer vision module

Adhering to the described system requirements, we implemented the computer vision module. Its structure is shown in fig. 5. The scheme is greatly simplified from the point of view of interaction with the knowledge base and reflects only the sequence and connectivity of computer vision modules. Let us consider these modules in more detail.

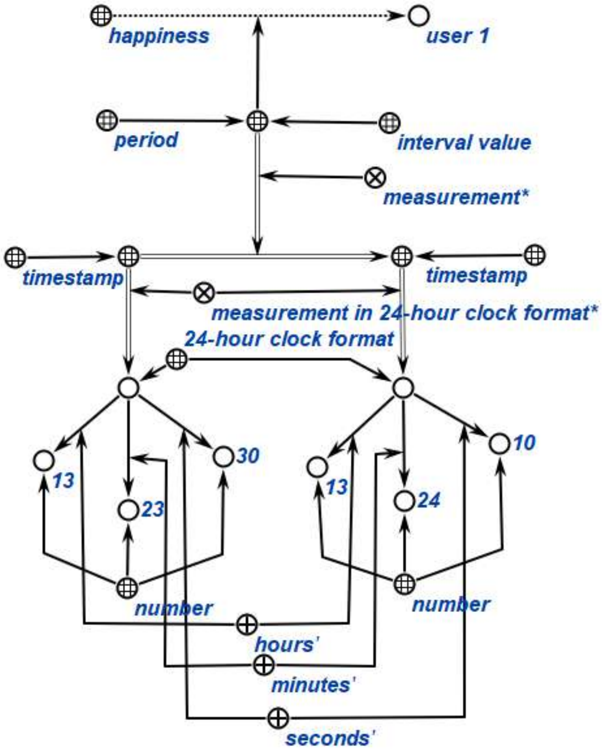

Figure 4. An example of a representation of the user and their emotions for a certain period in the knowledge base

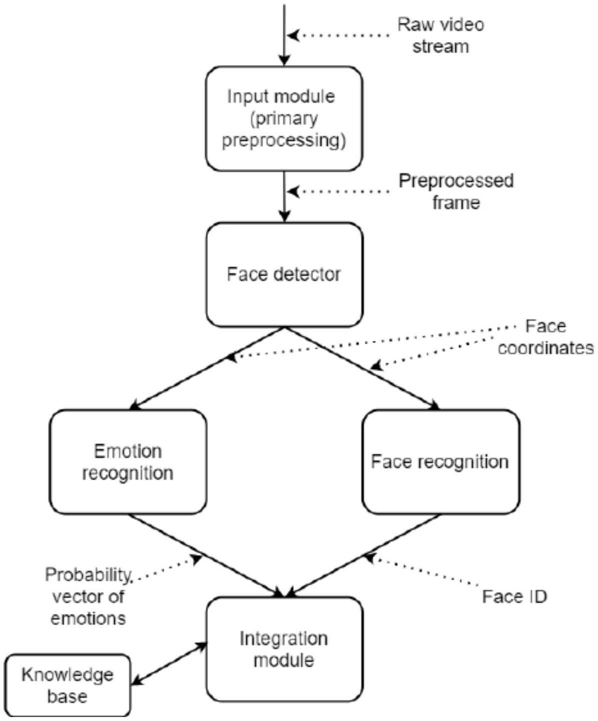

Figure 5. The scheme of the computer vision module pipeline

The **face detection module** solves the problem of detecting faces in the frame. It receives a frame from the video stream as input and returns the coordinates of the found faces.

The **identification module** is required to recognize the face of the user being identified. Receiving the coordinates of the detected persons as input, this module returns the user IDs.

The **emotion recognition module** is independent of the identification logic. Like the identification module, it receives the coordinates of the detected faces as input and returns a probability vector of emotions.

After the performance of particular branches of the pipeline, the results of the work of the modules are transmitted to the input of an integrator, which performs an immersion in the knowledge base.

Next, we will consider in more detail the main functions of the computer vision module and the results obtained in the preparation of appropriate neural network models.

B. *Functions of the computer vision module*

The functions of the computer vision module include:

- identification of a user known to the system;
- identification of unknown users after additional training;
- recognition of user emotions.

C. *User identification*

User identification is carried out by the identification module. The processes that occur in the module are most fully represented in the flowchart in fig. 6.

To implement this function, the FaceNet [12] neural network model is used. For this model, classical models of deep convolutional neural networks with a triplet loss function are used. In our case, the ResNet [13] convolutional network was used. The overall scheme of the model is shown in fig. 7.

This model consists of successive layers of a deep convolutional neural network and L_2 normalization using a triplet loss function at the training stage. At the output of the model, a 128-dimensional feature vector is formed, which can be used for the native comparison of faces.

The models from the dlib [14] library were used as the basic implementation.

Let us consider the user identification algorithm.

1) The user's face is detected by the detector in the frame (the MTCNN [15] model is used as a detector);
2) For the detected face, the feature vector using ANN is calculated;
3) After calculating the feature vector, it is compared with other feature vectors stored in the database. This comparison is performed with the preset threshold;

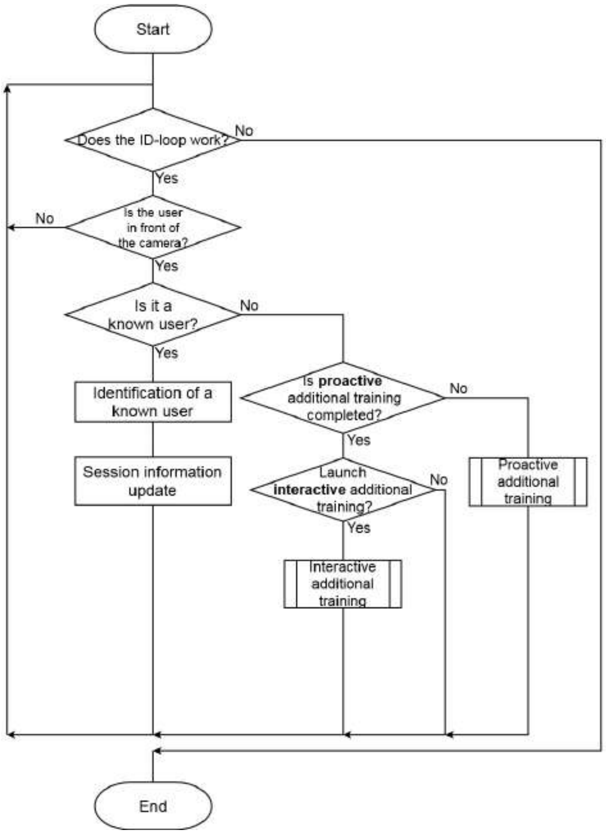
Start
Does the ID-loop work?
No
Yes
No
Is the user in front of the camera?
Yes
Is it a known user?
No
Yes
Identification of a known user
Is **proactive** additional training completed?
No
Yes
Session information update
Launch **interactive** additional training?
No
Proactive additional training
Yes
Interactive additional training
End

Figure 6. The overall flowchart of identification

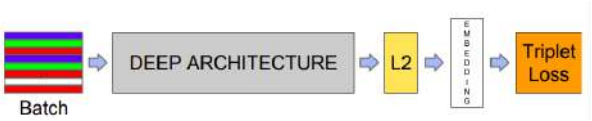

Figure 7. The structure of FaceNet (original image from [12])

4) Based on the results obtained in item 3, an ID is assigned to the detected face.

Thus, a necessary condition for the functioning of the algorithm is the presence of pre-calculated feature vectors for known users. This approach allows identifying the user with acceptable speed and accuracy.

The advantage of the proposed approach is that the implementation of recognition does not require a large training dataset of data, since the used FaceNet model is pre-trained on a large data array (a dataset of more than 3 million images) and can be used unchanged to identify people who were not included in this dataset.

For training, a dataset of photos of 7 different people – 7 photos of each person – was used. Thus, the dataset size was 49 photos. For testing, an independent test dataset of 14 photos and a set of video fragments were used to assess the quality of user recognition.

As a result of the evaluation of the proposed algorithm, the efficiency of face recognition was 95.84%. At the same time, the percentage of correctly recognized faces in the test dataset (due to its small size) was 100%.

D. Additional training for new unknown users

In addition to identifying known users by the feature vectors that are present in the database, the system allows performing additional training for recognizing unknown users in real-time. *Additional training* here means calculating a set of feature vectors for new users and saving them for further usage in the identification process. This process is carried out within so-called proactive and interactive additional training.

Proactive additional training (fig. 8) is conducted without direct user participation while calculating feature vectors based on frames received from the video stream.

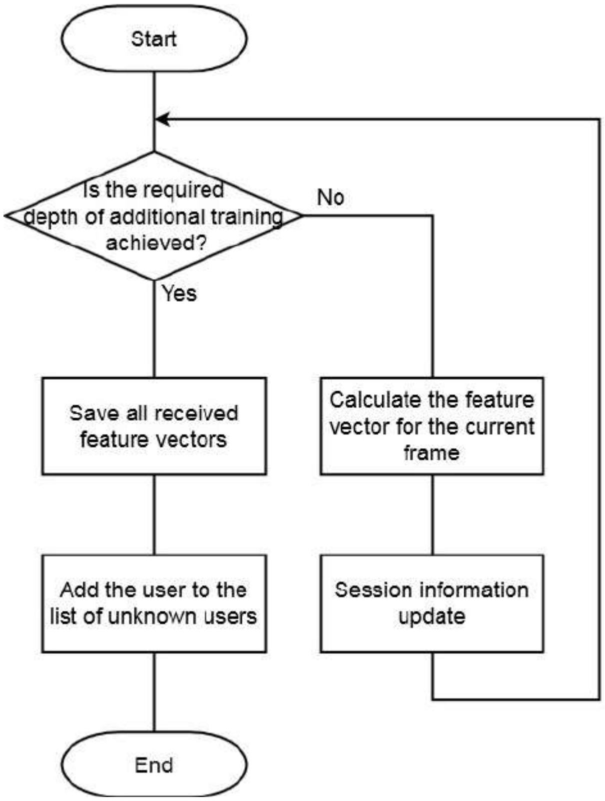
Start
Is the required depth of additional training achieved?
No
Yes
Save all received feature vectors
Calculate the feature vector for the current frame
Add the user to the list of unknown users
Session information update
End

Figure 8. The scheme of proactive additional training

Interactive additional training for new users (fig. 9) conducted separately and with the direct participation of the user allows improving the results of proactive additional training and getting more representative feature vectors.

E. Recognition of user emotions

The user emotions are recognized in seven basic classes: neutral facial expression, happiness, sadness, anger, fear, surprise and disgust.

The problem of classifying emotions by the image of the user's face is studied in many papers (for example, [16], [17], [18], etc.). With the active development of deep neural network training technologies, special emphasis by solving this problem began to be placed on the CNN-architectures (Convolutional Neural Networks) of various configurations. So, state-of-the-art results of emotion recognition were obtained for the eXnet [19] architecture. This network is used as a basic model in the proposed system. Its structure is shown in fig. 10.

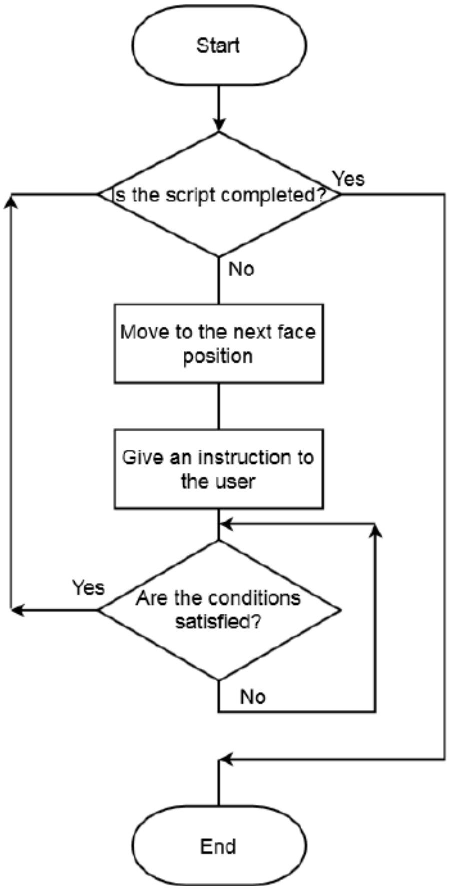
Start
Is the script completed?
Yes
No
Move to the next face position
Give an instruction to the user
Yes
Are the conditions satisfied?
No
End

Figure 9. The scheme of interactive additional training

In addition to choosing a model for recognizing emotions, the selection of the training dataset used is of great importance. So, in our work, a combined version of the training dataset from the well-known CK+ [20] dataset and datasets collected manually by the authors (the composition of the datasets is described in more detail below) is used. This approach allowed diversifying the training sample, making it less synthetic.

When forming the final dataset, images of faces with expressions of emotions from three basic sources were used:

1) The CK+ dataset. The training dataset consists of 4,615 images, the control dataset consists of 554 images structured according to 7 basic recognizable classes of emotions. This dataset consists of video fragments in the format of 640x490 or

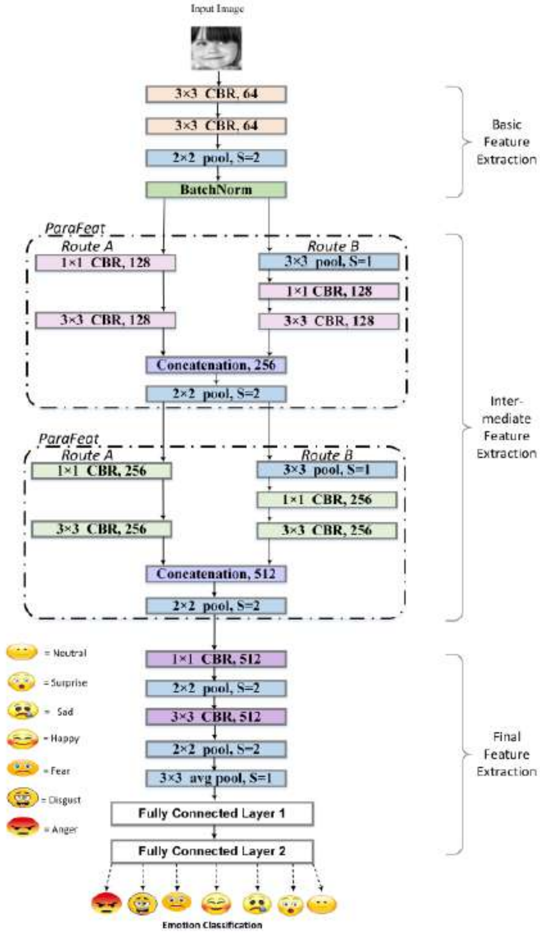

Figure 10. The structure of eXnet (original image from [19])

640x480 pixels. It should be noted, however, that the classes for this dataset differ from the above-mentioned ones (for example, instead of a neutral emotion, there is an emotion of despisal). This was one of the reasons why there was a need to form a combined dataset;

2) The ”Students and colleagues“ dataset. This dataset was collected from data obtained by self-determined image collection and processing. It is a set of images of the faces of 38 people with the expression of basic emotions. The training part included 20,947 images, the control part – 2,101;
3) Internet Emotions. This dataset was formed from images taken from the Internet and includes one-man photos of 295 people, which include training (268 images) and control (27 images) datasets.

Thus, the total size of the training dataset was 25,830 images and the control one – 2,682.

As a result of additional training of the eXnet model on the dataset described above, the results presented in table I were obtained.

Table I
Results of training the eXnet model

	avg. valid	*avg. valid softmax*
CK+	0.877	0.859
Students and Colleagues	0.765	0.742
Internet	0.37	0.305

avg valid – the percentage of successfully recognized emotions taken as 1 in the test dataset.

avg softmax – the percentage of successfully recognized emotions, taking into account the obtained probability from the range (0..1] in the test dataset.

ck+, **student_colleague** and **internet** denote the overall accuracy that results from the corresponding test datasets. **average_valid** and **average_valid_softmax** are the metrics used to evaluate accuracy.

F. Semantic analysis

The described integration mechanism allows enriching the knowledge base with the results of recognition of various models used by the computer vision module (identification model and emotion recognition model). The processing of this knowledge will be no different from the processing of any other knowledge in the ostis-system, regardless of whether they got there from the computer vision module, any sensors, visual or natural language interface or in some other way. In this case, computer vision is another receptor of the system.

Knowledge processing in the KB, i.e., semantic analysis, is performed by the problem solver. The problem solver is a set of agents that react to events in the knowledge base (for example, a problem definition), solve its problem (generating, transforming knowledge, accessing external systems) and put the result of the work in the same KB.

For example, one of the methods of knowledge processing can be the usage of logical inference [21], which generates new knowledge based on a set of rules. Logical rules, in the simplest case, can be represented by “if-then” bindings, where the “if” part describes the knowledge that must be in the knowledge base to make it possible for us to generate the knowledge described in the “then” part. The origin of such rules can be different: from adding them manually by knowledge base engineers to automatically generating them.

In the considered implementation of the hybrid system, the logical rules [21] are used to generate some standard system responses to the interlocutor’s messages. These rules use such knowledge as the identification of the interlocutor and their current emotion. Figure 11 shows a fragment of such a rule in a simplified form for clarity (in a real system, such rules have a more complex specification).

The meaning of the rule is as follows: if we received a greeting message from a user, whose emotion is recognized by the system as “sadness” and whose name the

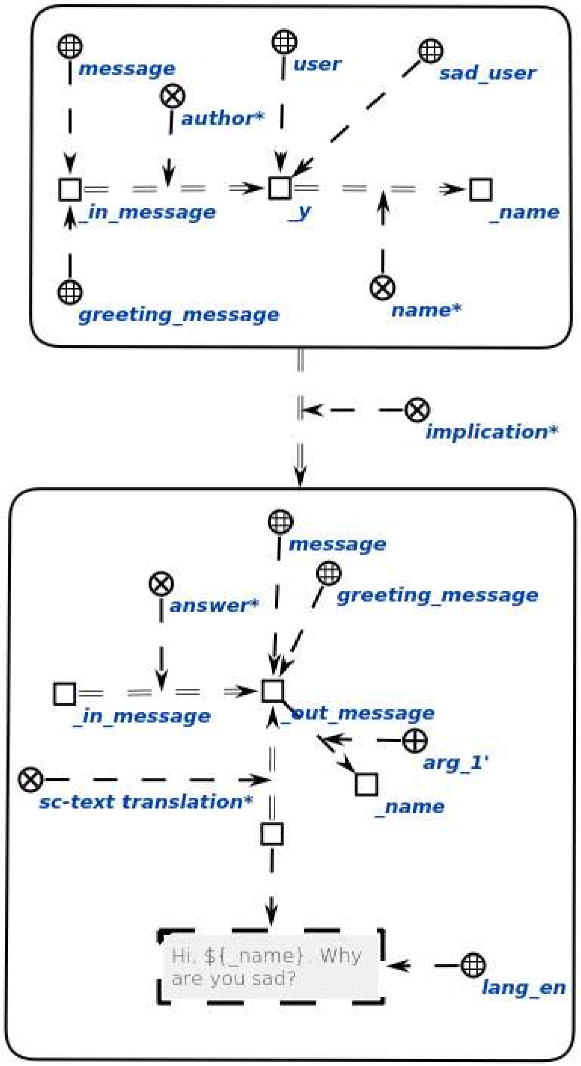

Figure 11. An example of a logical rule that uses the result of recognizing a user's emotion

system knows, then we need to respond to this message with a greeting with a reference by name and ask the reason for sadness.

V. Hardware architecture

The implementation of the represented models within the framework of a hybrid system for analyzing the emotional state requires appropriate support, both from the software and hardware. Therefore, the issue of creating a hardware architecture of the system that allows effectively implementing the functional responsibilities imposed on the system is one of the significant stages for achieving the overall goal.

The developed hardware architecture of the system should take into account the requirements and features of the implementation of both the semantic and neural network components of the system.

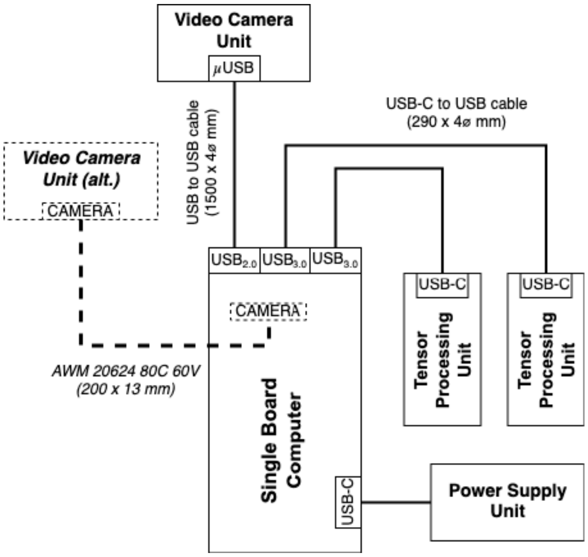

Figure 12. The system hardware architecture

The hardware requirements, from the point of view of the semantic part of the system, include the need to use computing tools with a processor based on the 'x86' architecture. This is due to the fact that initially the OSTIS platform as the basis of the software part of the system was developed for general-purpose CISC processors. Using this type of processor allows eliminating compatibility problems, simplifying debugging and testing the system. Therefore, it is necessary to have a computing device in the system based on this hardware architecture.

On the other hand, to solve computer vision problems, modern neural network architectures require support for tensor operations from the hardware platform, which allows effectively organizing the process of running trained neural network models on the target device. However, general-purpose processors are not suitable for performing such operations with maximum performance. Therefore, to increase the speed of the system, it is necessary to include a special coprocessor device in the hardware architecture, which allows increasing the speed of the neural network part of the system.

Taking into account the listed requirements, we have developed a hardware architecture of the system, the block diagram of which is shown in figure 12:

The main elements that build up the hardware of the system are:

- A single-board computer (SBC) that serves as a central device, on which the OSTIS virtual machine is run and, accordingly, the interpretation of intelligent agents and a list of peripheral devices that perform the functions of input and output of video and audio information as well as auxiliary devices that perform the functions of supporting neural network computing is carried out;

- To input video information, a camera (Video Camera Unit – VCU) is used designed to solve computer vision problems: detecting and tracking the user's face in the frame, identifying the user, recognizing the emotions of the subject in the frame based on frames from the video stream. According to the general scheme of the system model 1, the camera transmits the video stream to the computer vision module (computer vision pipeline), and only then the information is sent to the internal part of the system for further processing. The peculiarity of the proposed solution, which gives it additional flexibility, is that the computer vision module can be deployed both on the single-board computer itself and on a separate machine (for example, a PC or a specialized device designed to solve problems of this type). Therefore, the camera can be connected both to a single-board computer directly and to an external device that will transmit information through the abovementioned module to the central device where the OSTIS system core is run. The scheme of the hardware architecture considers the option when the camera is connected directly;
- To speed up the performance of operations on vectors and matrices when calculating neural networks on hardware with limited resources, which include single-board computers, it is proposed to use tensor coprocessors (Tensor Processing Unit – TPU) for neural network calculations.

Let us focus on each of the hardware components of the system and consider them in more detail.

A. *Single-board computer*

A single-board computer (SBC) is a computer set up on a single printed circuit board, on which a microprocessor, RAM, I/O systems and other modules necessary for the operation of a computer are installed [22], [23].

As the basis of the hardware platform, it was proposed to use a single-board computer, since such a form factor, on the one hand, can provide the necessary and sufficient runtime environment for the OSTIS Technology in terms of performance and functionality and, on the other hand, preserve the minimum weight-and-dimensional and cost characteristics of computing tools, including for solving problems connected with the Internet of things and the usage of OSTIS within the framework of the "Edge Computing and AI" [24], [25] concept.

We have reviewed and compared the models of single-board computers on the local and international markets that are available for delivery to the territory of the Republic of Belarus. The model lines of computers from such manufacturers as Interl NUX, LattePanda, Rock Pi, UDOO, ODYSSEY [26] were considered. We have set the maximum cost of a single-board computer, so that it does not significantly exceed the cost of the option for the ARM architecture. The set cost was no more than $100, which significantly narrowed the search area.

Figure 13. A single-board computer "Rock PI X Model B"

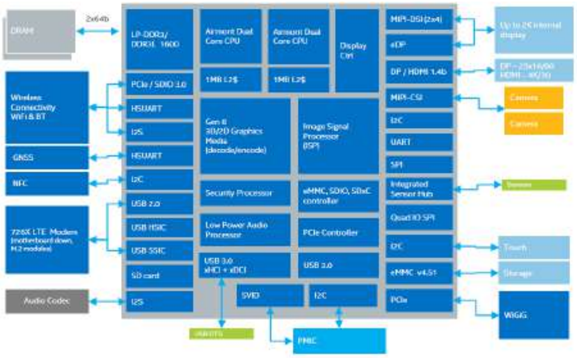

Figure 14. A functional diagram of the Intel Atom X5 processor [28], [29]

Among the currently available models of single-board computers, the Rock Pi computer has become the most preferred option, namely the Rock PI X Model B model [27], the appearance of which is shown in fig. 13.

The functional diagram of the CPU [28], [29] is shown in figure 14.

A distinctive feature of this single-board computer is the presence of ROM based on eMMC, which allows ensuring the functioning of the system and high-speed access to data on a solid-state drive without using an external SD drive. The maximum available capacity is 128 GB. The proposed architecture uses a version with 32 GB of memory, which is sufficient to contain the OS as well as the necessary software modules and OSTIS intelligent agents.

B. *Video camera*

It is an element of the system, through which a video stream is received and transmitted to a single-board computer to solve subproblems connected with computer vision and recognition of visual images. It acts as the

Figure 15. The Sony PS3 Eye camera

Figure 16. The Coral USB Accelerator TPU device

main channel of input information for the neural network modules of the system.

Within the current version of the hardware architecture, it is possible to connect various video cameras, depending on the type of video interface and the configuration of the device package, which will be directly determined by the type of problem being solved and the requirements for viewing angles, focal point, depth of field of the shown space, dimensions and scale of recognized objects in the image. The main difference between them is the design of the camera box itself as well as the type of interface that is used to connect them.

The camera is a separate device in its box, which can be set on a tripod and shoot at the level of the user's face. As such a removable camera, it is proposed to use the Sony PS3 Eye device (fig. 15) [30], [31].

It should be noted that both cameras are used in a mode of obtaining images of relatively low resolution, i.e., 640 x 480 60 fps, 320 x 240 120 fps. This is done for two purposes: the first is to increase the performance of the video subsystem as well as to prevent the image obtaining process, which is resource-consuming for single-board computer processors, from becoming a "bottleneck" in the common processing pipeline; the second is that the image will be transformed into images with a lower resolution one way or another before processing by neural network models. Such a transformation is called "oversampling", or "resampling", that is usually performed in all machine vision systems to ensure a balance between performance and quality of work, since processing large-resolution images requires significantly large computing resources as well as time for training and performing neural network models. In the case of our proposed system, image resampling is carried out for user identification models in the format of 160x160 pixels as well as in the format of 48x48 pixels – for emotion recognition models. For this reason, for the development of the hardware architecture, we chose the Sony PS3 Eye camera, which is significantly inferior in characteristics to modern cameras but, on the other hand, has a minimum cost on the market relative to the quality of the optical system and the CCD-matrix installed in it, which allow solving the entire range of denoted machine vision problems.

C. Neural network computing accelerator

An important element of the hardware architecture is a tensor coprocessor for neural network computing to speed up the work of the neural network modules of the application [32].

We have considered the model lines of tensor coprocessors of the main manufacturers of specialized purpose processors from Intel, NVidia and Google [33], [34]. The most suitable option in terms of performance characteristics, cost, amount of documentation and open source projects is the "Coral USB Accelerator" processor, which was chosen as this component of the system [35].

This processor was developed by the Google corporation and is intended for usage with the TensorFlow machine learning library. This device, in comparison with GPUs, is designed to perform a large number of calculations with reduced accuracy (in integer arithmetic) with higher performance per watt. The device is implemented as a matrix multiplier controlled by the instructions of the central processor over the USB 3.0 bus.

The Coral USB Accelerator coprocessor can perform 4 trillion operations (teraflops) per second (TOPS), using 0.5 watts for each TOPS (2 TOPS per watt). For example, it can perform tensor calculations for one of the most popular neural network architectures in technical vision problems, such as "MobileNet v2", with a performance close to 400 frames per second with a low energy consumption of about 1.5 W [36].

CONCLUSION

The proposed model of neuro-symbolic AI is an example of combining different directions of AI. This model allows using various neural network models to solve computer vision problems and conduct a semantic analysis of the results of the work of such models in the knowledge base. It is also possible to add new neural network models and control their operation mode through the knowledge base.

The proposed model is used to implement a hybrid system of semantic analysis of the emotional state of the user, which operates with knowledge formed in the process of interaction with neural network models.

In the article, a variant of a hardware platform for the operation of the developed system based on single-board computers of the "Raspberry Pi 4B" and "Rocks Pi X" models as well as "Raspberry Pi Camera v1.3" and "Sony PS3 Eye" video cameras and the "Google Coral USB Accelerator" tensor processor is also proposed.

The proposed hardware and software architecture provides the necessary level of performance and mobility for the semantic and neural network parts of the system.

The described model creates the basis for further research in the field of developing:

- universal integration with the knowledge base of any neural network models (not only solve computer vision problems);
- an approach to deeper integration of neural network models with the knowledge base, when through the knowledge base it becomes possible to control not only the operating mode of neural network models but also their topology, architecture, combination with other models, etc.;
- an approach to automatic decision-making on the usage of a particular neural network model for solving system problems;
- an approach to usage of the knowledge base to improve the training of artificial neural networks;
- new hardware architectures that can support such systems.

Acknowledgment

The authors would like to thank the research groups of the Departments of Intelligent Information Technologies of the Belarusian State University of Informatics and Radioelectronics and the Brest State Technical University for their help and valuable comments. In particular, Vladimir Adamovich Golovko, Vladimir Vasilyevich Golenkov, Daniil Vyacheslavovich Shunkevich. We also thank Intelligent Semantic Systems LLC for providing the equipment for conducting research.

References

[1] V. Golovko, A. Kroshchanka, M. Kovalev, V. Taberko, and D. Ivaniuk Neuro-Symbolic Artificial Intelligence: Application for Control the Quality of Product Labeling, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2020, pp. 81–101.

[2] Castelvecchi D. (2016). Can we open the black box of AI?. Nature, 538(7623), pp. 20–23. https://doi.org/10.1038/538020a

[3] J. Sowa Semantic networks, in encyclopedia of artificial intelligence, Expert Systems with Applications (1987)

[4] F. Lehmann Semantic networks, Computers Math. Applic 23(2–5), 1–50 (1992)

[5] T. Besold, A. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K. Kuehnberger, L. Lamb, D. Lowd, P. Lima, L. de Penning, G. Pinkas, H. Poon, G. Zaverucha Neural-symbolic learning and reasoning: A survey and interpretation, (Nov 2017), https://arxiv.org/pdf/1711.03902.pdf, (accessed 2021, Jun)

[6] Garcez, A., Besold, T. R., Raedt, L., Foldiak, P., Hitzler, P., Icard, T., Kuhnberger, K-U., Lamb, L. C., Miikkulainen, R. and Silver, D. L. Neural-Symbolic Learning and Reasoning: Contributions and Challenges, AAAI Spring Symposium Series, 23-03-2015

[7] A. d'Avila Garcez, L. Lamb, D. Gabbay Neural-symbolic cognitive reasoning, In: Cognitive Technologies, Springer (2009)

[8] B. Hammer, P. Hitzler Perspectives of neural-symbolic integration, In: Studies in Computational Intelligence (77), Springer (2007)

[9] V. Golovko, V. Golenkov, V. Ivashenko, V. Taberko, S. Ivaniuk, A. Kroshchanka, M. Kovalev Integration of artificial neural networks and knowledge bases, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2018, pp. 133–145

[10] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, Methodsand tools for ensuring compatibility of computer systems, Otkrytyesemanticheskie tekhnologii proektirovaniya intellektual'nykh system[Open semantic technologies for intelligent systems], 2019, pp. 25–52.

[11] Intelligent meta system(IMS). Available at: http://ims.ostis.net (accessed 2021, Jun)

[12] F. Schroff, D. Kalenichenko, and J. Philbin FaceNet: A Unified Embedding for Face Recognition and Clustering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[13] K. He, X. Zhang, S. Ren, and J. Sun Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[14] D. E. King Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research, vol. 10, 2009, pp. 1755–1758.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks, in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[16] C. F. Benitez-Quiroz, R. Srinivasan and A. M. Martinez EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5562–5570.

[17] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition, in arXiv, arXiv:1905.04075v2 [cs.CV], https://arxiv.org/abs/1905.04075, 2019.

[18] S. Shojaeilangari, W. Yau, K. Nandakumar, J. Li, and E. K. Teoh Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning, in IEEE Transactions on Image Processing, vol. 24, no. 7, pp. 2140–2152, July 2015.

[19] M. N. Riaz, Y. Shen, M. Sohail, and M. Guo eXnet: An Efficient Approach for Emotion Recognition in the Wild, in Sensors, vol. 20(4), 2020.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops, 2010, pp. 94–101.

[21] V. Vagin, A. Zagoryanskaya, M. Fomina Dostovernii i pravdopodobnii vivod v intellektualnih sistemah, Moscow: FIZMATLIT, (in Russian) p. 704 p. (2008)

[22] Single-Board Computer. Available at: https://en.wikipedia.org/wiki/Single-board_computer (accessed 2021, May)

[23] U. Isikdag Internet of Things: Single-board computers, Enhanced Building Information Models, Springer Cham, 2015. pp. 43–53.

[24] P. Li Arhitektura interneta veshhej, DMK Press (in Russian), 2018, 454 p.
[25] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, & X. Chen Edge AI: Convergence of Edge Computing and Artificial Intelligence, Springer Nature, 2020, 218 p.
[26] P. Galkin, L.Golovkina, I. Klyuchnyk Analysis of single-board computers for IoT and IIoT solutions in embedded control systems, 2018 International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), IEEE, 2018, pp. 297–302.
[27] ROCK Pi X Hardware Information. Available at: https://wiki.radxa.com/RockpiX/hardware (accessed 2021, May)
[28] Intel Atom Z8000 Processor Series Vol. 1. Available at: https://dl.radxa.com/rockpix/docs/hw/atom-z8000-datasheet-vol-1.pdf (accessed 2021, May)
[29] Intel Atom Z8000 Processor Series Vol. 2. Available at: https://dl.radxa.com/rockpix/docs/hw/atom-z8000-datasheet-vol-2.pdf (accessed 2021, May)
[30] Play Station Eye. Available at: https://en.wikipedia.org/wiki/PlayStation_Eye (accessed 2021, May)
[31] PS3 Eye Camera Technical Specification. Available at: https://icecat.biz/en/p/sony/9473459/webcams-eye-camera-+ps3-1269549.html (accessed 2021, May)
[32] AI accelerator. Available at: https://en.wikipedia.org/wiki/AI_accelerator (accessed 2021, May)
[33] Y. Chen, Y. Xie, L. Song, F. Chen, T. Tang A survey of accelerator architectures for deep neural networks, Engineering, vol. 6 no. 3, 2020, pp. 264–274.
[34] W. Li, M. Liewig et al. A Survey of AI Accelerators for Edge Environment, Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing, vol 1160. Springer, Cham, 2020, pp 35–44.
[35] Coral USB Accelerator Datasheet. Available at: https://coral.ai/docs/accelerator/datasheet (accessed 2021, May)
[36] Edge TPU Performance Benchmark. Available at: https://coral.ai/docs/edgetpu/benchmarks (accessed 2021, May)

Семантический анализ видео-потока на основании нейро-символического искусственного интеллекта

Крощенко А.А., Михно Е.В.,
Ковалев М.В., Захарьев В.А., Загорский А.Г.

Статья посвящена модели компьютерного зрения, базирующейся на нейро-символическом подходе. Приведена архитектура предлагаемой модели с подробным описанием составляющих ее компонентов. Описаны основные применения преимущества подобной модели на примере диалоговых систем. Во второй части работы приводится пример разработки системы-компонента диалоговой системы для оценки эмоционального состояния пользователя, базирующейся на предложенной нейро-символической модели. Показано, что класс подобных систем сочетает в себе преимущества коннекционистского и символического подхода в искусственном интеллекте. Приводится обзор аппаратной платформы, позволяющей осуществлять запуск и поддержку работы системы в компактном форм-факторе одноплатного компьютера.

Received 09.07.2021