Design of a Program Complex Architecture for Equipment Control Using Gestures

Mikhail D. Artemov[1], Lilia I. Voronova[2]
Department of Intelligent Control Systems and Automation
Moscow Technical University of Communication and Informatics
Moscow, Russia
[1]artemov_mikle@mail.ru, [2]voronova.lilia@yandex.ru

***Abstract*. Industrial Internet of Things technologies is widely distributed. Smart cities, factories, shops, house assume using intelligent methods of data processing providing by a sensors. In particular, smart enterprise control is possible by through of gestures and is interesting and perspective task. In the article briefly reviewed methodology of gesture and action recognition by image sequences and proposed neural network architecture for gesture recognition and smart enterprise control.**

***Key words:* gesture recognition, Mask R-CNN, CNN, 3D CNN, LSTM.**

I. INTRODUCTION

Industrial Internet of Things (Internet of Things for industrial/corporate use, IIot) is system of united computer networks and connected industrial facilities having an integral sensors and a software for data collection and exchange, which allow remote control and management in automated mode without a human control.

IIoT has the following dignities:

- Paper documentation is can being rejected.
- Specific knowledge of a specialists can be collected.
- Unscheduled delay, maintenance and equipment breaking are being prevented and fail of delivery sequence can be delayed. Consequently, efficiency of enterprise work can been improved.
- Production stages can be automated and optimized.
- Introduction of network interaction between a machine, an equipment, knowledge and information systems allow effective monitoring and analysis of environment, production process and real-time self-state.
- Risk of emergency can be decreased due to prevent of malfunction and decreasing risk of manufacture delays.

There are the following principles: firstly, sensors, actuating mechanisms, controllers and man-machine interfaces are installed to important equipment parts. Secondly, information is collected and allow to get objective and accurate data. Handled data move to all the enterprise departments and allow to accept a substantiate decisions and to establish interaction between a different departments employees.

For implement of this approach, all necessary information about real resources state within the enterprise and within other enterprises is to be made available to automated control systems of different levels.

Introduction of IoT requires to change approaches of create and use of automated information control systems and common approaches of enterprises control [1]. Obsolete production lines, which cannot be automated using IoT will be changed to new automate and a robotic equipment in future.

For traditional enterprises and them systems, a staff is the base resource, which is necessary for other kinds of resource control. Consequently, voice information exchange and data between employees is main way of information exchange in such systems.

Grow and development of traditional enterprises is related to using IIoT. Adequate interpretation and filtration is high priority task for enterprises, when they are processing huge quantity of unstructured data, and correct information presentation understandable for user is a task acquiring special mean. For those tasks, a modern market is offering an advanced analytical platforms designed to combine, store a analyze data about technical processes and events in real time.

We offer to unify man-machine interface for interaction with equipment; this will allow to simplify handle of huge quantity of data using single information representation and processing.

Unification is meant a control using gestures. Gest is some action or motion of a human body or a him part, which has special meaning, i.e. is sign or symbol [2]. Reject of traditional mechanisms in favor of neural network and a single video camera explicit advantage such control [2; 3].

The video used for control is valuable data source for IoT. So, for example, using machine learning technology will allow to automate processes of algorithms improving, which are being executed by programming to “management cloud”, i.e. this allows to optimize management algorithms, when historical data are being combed, and, consequently, to increase a management efficiency. Collected data can train neural networks to action algorithms, which an operator executes, and in case of unusual system behavior can to notify about that. Moreover, trained system will give remote control opportunity or can independently to launch studied action algorithms [4; 5].

The single interface will allow to unify support such system and to run control of operators efficiency work on a different workstations.

Such system could to provide security of different levels by changing authentication politics and using artificial intelligence, when it analyzes a video frame. Moreover, such interface will be needed single knowledge and skills in the field of cybersecurity, this will allow to simplify them in future. Moreover, single system will ensure data compatibility and give opportunity of secure automated system of patch installing.

New technologies allow enterprises of different industries to achieve significant competitive advantages and take steps to meet recommendations for develop and exploit of IoT.

Such decisions transform production data to useful information, which is required for security and rational enterprise management.

II. RELATED WORK

Existing gesture recognition networks is can categorized by a principle of work with time dimension into different three groups, as proposed in [6]:

- Networks are using 3D filters in the convolutional layer (Fig. 1, *a*). The 3D convolution and 3D pooling in CNN layers allow to capture discriminative features along both spatial and temporal dimensions while maintaining a certain temporal structure.
- Networks are using motion features like 2D dense optical flow maps, which can be precomputed and input to the networks (Fig 1, *b*). Extracted motion features can be fed to the network as additional channels to the appearance ones or input to a secondary network (later combined with
- Networks are using combines a 2D (or 3D) CNN applied at individual (or stacks of) frames with a temporal sequence modeling (Fig. 1, *c*). Recurrent Neural Network (RNN) is one of the most used networks for this task, which can take into account the temporal data using recurrent connections in hidden layers.

3D filters in the convolutional layer are used [7; 8]. A space-temporal feature learning approach is offered [7], which use a deep 3D convolutional networks, which are trained on large number controlled video datasets. Their findings are three-fold:

- they are more suitable for spatiotemporal feature learning compared to 2D ConvNets;
- a homogeneous architecture with small convolution kernels in all layers is among the best performing architectures for 3D ConvNets;
- a simple linear classifier outperforms state-of-the-art methods and achieving 52.8% accuracy on UCF101 dataset with only 10 dimensions and is also very efficient to compute due to the fast inference of ConvNets.

This network conceptually very simple and easy to train and use.

It is proposed hand gesture recognition system that interleaves depth and intensity channels to build normalized spatio-temporal volumes, and train two separate subnetworks with these volumes [8]. The VIVA challenge's hand gesture dataset contains 885 intensity and depth video sequences of 19 different dynamic hand gestures performed by 8 subjects inside a vehicle. Both channels were recorded with the Microsoft Kinect device and have a resolution of 115×250 pixels.

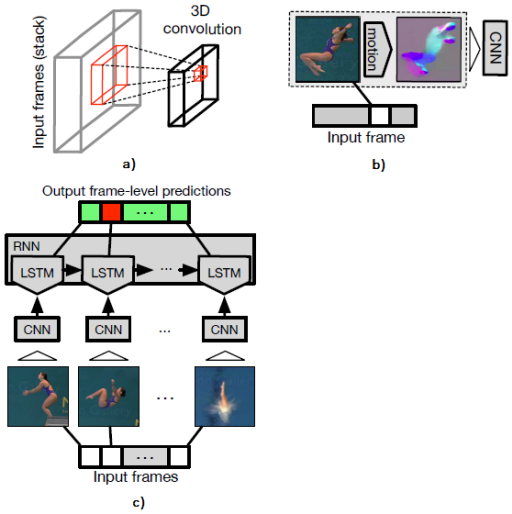

Fig 1. Groups of neural networks:
a) Networks are using 3D filters in the convolutional layer; b) Networks are using motion features like 2D dense optical flow maps; c) Networks are using combines a 2D (or 3D) CNN applied at individual (or stacks of) frames with a temporal sequence modeling [6]

An approach [9] is related to the second group propose semi-supervised approach using a deep neural network, by combining an autoencoder with a classification loss function, and training both of them in parallel. Other work [10] is related to same group propose three simple, compact yet effective representations of depth sequences, referred to respectively as Dynamic Depth Images (DDI), Dynamic Depth Normal Images (DDNI) and Dynamic Depth Motion Normal Images (DDMNI).

Other approach [11] is related to third group, which use a very large, annotated video dataset of the dynamic hand gestures and neural networks are trained on this data. This solution require a single camera and worked on various of platform. To train system, used a large dataset of short, densely labeled video clips that was crowd-acted by our community of crowd workers. The dataset contains ~150,000 videos across 25 different classes of human hand gestures, split in the ratio of 8:1:1 for train/dev/test. It also includes two "no gesture" classes to help the network distinguish between specific gestures and unknown hand movements.

Long Short-Term Memory Recurrent Neural Network (CNNLSTM) is proposed for the problem of dynamic gesture recognition[12]. That model consists of two consecutive convolutional layers, a flattening layer, a Long Short-Term Memory recurrent layer and a softmax output layer.

III. MASKCNNLSTM

In our project, we use approach based on using of a 3D CNN, which is individually applied to set of frames, which are modeling a temporal sequence.

Gesture control requires user's gesture recognition in real-rime. Therefore, the system should divide the video sequence into frames at a speed of 15–20 frames per second.

Task target neural network is human gestures recognition. Therefore, on introductory stage, we should detach a human and

his hands. This problem is solved by segmentation using by a Mask R-CNN (Fig. 2).

Apart from a good results of instance segmentation and object detection,Mask R-CNN is suitable for determining human pose estimation in photographs. For this, the keypoints select is important. That is a left shoulder, a right elbow, a right knee, etc. Using such points allows you to draw human pose skeleton. A neural network is being trained for select of the keypoints, and after that it should receive masks, in which just a one pixel have 1, and all other have 0. At the same time, the network is being trained to output the several single-pixel masks one at each keypoint [13].

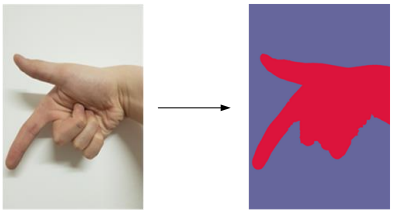

Fig. 2. The input image and the Mask R-CNN result

Use of segmentation will allow to simplify training of neural network, which is being making gesture recognition. Moreover, this lets to increase a classification accuracy and to detail human pose estimation with the aim of improve gesture definition, when key points are being added.

For recognition we propose to use MaskCNNLSTM architecture, which shows in Fig. 3.

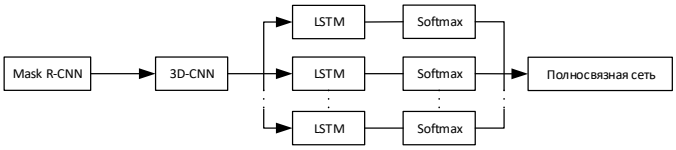

Fig. 3. Proposed gesture recognition architecture (MaskCNNLSTM)

After segmentation the frames sequence, on first stage, input to the 3D-CNN. That network handles the frames as seen in Fig. 4.

1) An input is being got sequence of the L frames.

2) Convolutional layers have filters of size [*k*,*k*,*d*], there *k* is dimension of reception field; *d* is frames count, which will be handled by filter. In our case, *d* = *L*, i.e the filter have [*k*,*k*,*L*] size. Our network will have the small height and weight of a convolution kernel (for example, 2×2 or 3×3) with the same filter size on all convolutional layers, since this will allow to receive best performing architecture [7].

3) In the output of the network we get a 3D feature map of size [*f*,*h*,*n*], which can be considered as sequence of *n* map with size [*w*,*h*].

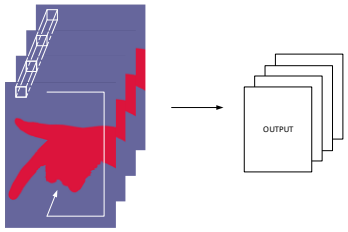

Fig. 4. Handling of the image sequence using the 3D-CNN

Features, which got from the 3D-CNN and related with action, can exist only at special position or time periods. Consequently, it is necessary to introduce a focusing mechanism for improve classification efficiency. Therefore, the feature maps received from the 3D-CNN are entering to inputs of LSTM networks, which can naturally introduce a temporal memory using previously received context.

The LSTM has been successfully combined with 2D CNN to incorporate visual attention and locate region of interest in the video sequence.

Usually, the LSTM incorporates all the input as a vector, which could not preserve the spatial correlation within an image. For solution of that problem, we will use a model [14]. The model is based on the LSTM, which combines the temporal and spatial attention mechanisms into a single scheme.

The LSTM model is being described by following equations:

$$i_t = \sigma\left(x_t W_{xi} + h_{t-1} W_{hi} + c_{t-1} W_{ci} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(x_t W_{xf} + h_{t-1} W_{hf} + c_{t-1} W_{cf} + b_f\right) \quad (2)$$

$$o_t = \sigma\left(x_t W_{xo} + h_{t-1} W_{ho} + c_{t-1} W_{co} + b_o\right) \quad (3)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left(x_t W_{xc} + h_{t-1} W_{hc} + b_c\right) \quad (4)$$

$$h_t = o_t \circ \tanh\left(c_t\right), \quad (5)$$

where x_t is the input to the LSTM block; i_t, f_t, o_t, c_t, h_t are the input gate, the forget gate, the output gate, the cell state and the output of the LSTM block respectively at the current time step *t*. *L* is the weights between the input layer and the input gate, the forget gate and the output gate respectively. W_{hf}, W_{hi}, W_{ho} are the weights between the hidden recurrent layer and the forget gate, the input gate and the output gate of the memory block respectively. W_{ci}, W_{cf}, W_{co} are the weights between the cell state and the input gate, the forget gate and the output gate respectively and finally; b_i, b_f, b_o are the additive biases of the input gate, the forget gate and the output gate respectively. The set of activation functions consists of the sigmoid function $\sigma()$, the element-wise multiplication $\circ()$ and the hyperbolic activation function $\tanh()$.

Feature map was weighted by a spatial attention weight matrix, this allows to safe spatial attention.

$$\alpha_{t,i} = \frac{\exp\left(W_i e_{t,i}\right)}{\sum_{j=1}^{w \times h} \exp\left(W_i e_{t,i}\right)}, \quad (6)$$

where $X_{t,i}$ is feature map [*m*,*h*] received from 3D-CNN; *i* is cell of $X_{t,i}$; W_i is the weight mapping to element *i* of the spatial weight matrix $X_{t,i}$; $e_{t,i}$ is multilayer perceptron conditioned on the current input $X_{t,i}$ and the previous hidden state h_{t-1}:

$$e_{t,i} = \tanh\left(W_{\alpha} X_{t,i} + W_{\alpha h} h_{t-1} + b_i\right). \quad (7)$$

The $\alpha_{t,i}$ evaluates the importance of the i region to the frame per point t.

$$x_t = \sum_{i=1}^{w \times h} \alpha_{t,i} X_{t,i} \ . \quad (8)$$

Thus, at each time step, the LSTM will predict the weight matrix for the next time step, and output of the LSTM will includes two parts, which are the activity labels and the weight matrix.

Every a softmax layer makes a predict for each the LSTM output. Finally, a fully connected network determines class of the source videosequence.

IV. CONCLUSION

Recently, interest in recognizing actions and gestures has much increased. In the this work, we briefly review a methodology of gesture recognition by an image sequences and propose a neural network architecture MaskCNNLSTM, which include a Mask R-CNN, a 3D CNN and a LSTM networks, which will be to used for gesture recognition.

A proposed system will could to provide security of different levels by changing authentication politics and using artificial intelligence, when analyze a video frame. Moreover, such interface will be requiring single knowledge and skills in the field of cybersecurity, this will allow to simplify them in future. Moreover, single system will ensure data compatibility and give opportunity of secure automated system of patch installing.

V. REFERENCES

[1] Воронов В.И., Усачев В.А. Компетенция «машинное обучение и большие данные» // Приоритетные направления развития науки и образования. Под общ. ред Г.Ю. Гуляева. Пенза, 2017. С. 97–108. [Usachev V.A., Voronov V.I. The competence “Machine learning and Big Data” // Priorities for development of science and education / Ed. G.Yu. Gulayev. Pensa, 2017. P. 97–108].

[1] Михаеску С.В., Трунов А.С., Воронова Л.И. Анализ предметной области для разработки системы построения скелетной модели человека на основе массива опорных точек, получаемых совокупностью контроллеров Kinect // Международный студенческий научный вестник. 2015. № 3-4. С. 521–522. [Mihaesku S.V., Trunov A.S., Voronova L.I. Analysis of the subject area of system developing for building a human skeletal model based on an array of reference points obtained from a set of Kinect controllers // International Student Scientific Bulletin, 2015. N 3-4. P. 521–522].

[2] Воронов В.И., Воронова Л.И., Генчель К.В. Применение параллельных алгоритмов в нейронной сети для распознавания жестового языка // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2018). VII Международная научно-техническая и научно-методическая конференция: Сб. науч. ст.: В 4 т. / Под ред. С.В. Бачевского. М., 2018. С. 207–212. [Voronov V.I., Voronova L.I., Genchel K.V. The use of parallel algorithms in the neural network to recognize the sign language // Actual problems of information and telecommunications in science and education (APINO 2018). VII International scientific tech. and scientific method. Conf.: In 4 t. / Ed. S.V. Bachevsky. M., 2018. P. 207–212].

[3] Горячев Д.В., Воронов В.И. Большие данные и машинное обучение // Технологии информационного общества: Матер. XII Междунар. отраслевой науч.-техн конф. М.,2018. С. 327–328. [Goryachev D.V. Big Data and Machine Learning // Information Society Technologies: Proceedings of 12th the international scientific-practical conference. Moscow, 2018. P. 327–328].

[4] Voronov V.I., Voronova L.I. Features of realization master’s program “Automation of technological processes and manufactures” // International Journal of Applied and Fundamental Research. 2016. № 2. URL: www.science-sd.com/464-25196.

[5] Asadi-Aghbolaghi M., Clapes A., Bellantonio M. et al. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences // 12th IEEE Conference on Automatic Face and Gesture Recognition. 2017. P. 476–483.

[6] Tran D., Bourdev L., Fergus R. et al. Learning Spatiotemporal Features with 3D Convolutional Networks // IEEE International Conference on Computer Vision. 2015.

[7] Molchanov P., Gupta Sh., Kihwan Kim K. et al. Hand Gesture Recognition with 3D Convolutional Neural Networks // IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015.

[8] Gupta O., Raviv D., RaskarR. Multi-velocity neural networks for gesture recognition in videos. URL: https://arxiv.org/abs/1603.06829.

[9] Wang P., Li W., Liu S. et al. Large-scale Isolated Gesture Recognition Using Convolutional Neural Networks // 23rd International Conference on Pattern Recognition (ICPR). Cancun, 2017.

[10] Gesture recognition using end-to-end learning from a large video database. 2017. URL: https://medium.com/twentybn/gesture-recognition-using-end-to-end-learning-from-a-large-video-database-2ecbfb4659ff.

[11] Tsironi E., Barros P., Wermter S. Gesture Recognition with a Convolutional Long Short-Term Memory Recurrent Neural Network // Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). 2016. P. 213–218.

[12] MaskR-CNN: Architecture of modern neural network for object segmentation by image. 2018. URL: https://habrahabr.info/development/image-processing/4391-mask-r-cnn-architecture-of-a-modern-neural-network-for-segmenting-objects-on-images.html.

[13] Lu N., Wu Y., Feng L. et al. Deep Learning for Fall Detection: Three-Dimensional CNN Combined With LSTM on Video Kinematic Data // IEEE Journal of Biomedical and Health Informatics. 2019. Vol. 23, N 1. P. 314–323.