Gesture recognition system based on Convolutional neural networks

I S Chistyakov* and E V Chepin
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, 115409 Moscow, Russian Federation

*Corresponding author’s e-mail address: chist-34dml@mail.ru

Abstract. This article tells about gesture recognition system based on Convolutional neural nets. System consists of two parts: the tracking and detection subsystem and gesture recognition subsystem.

1. Introduction.
At the moment, the research of interaction between human and technical devices has received growing attention by the development technologies. Development natural ways for human and machine is actual task. It makes communication easier and more effective and expands the area of problems to solved the computer.

We have several tools for interaction: voice, facial expressions, gestures etc. One of management variants of robotic objects is static and dynamic hand gestures.

Static gestures is some position that can be fixed (Fig. 1).

Dynamic gesture is gesture that performs for period of time (Fig. 2)
Gesture - econtinuous action, so we can’t to determine what kind of gesture was performed basing only on current object position. We must store passed object positions and check this array.

Quality system requires reliable performance a several subsystems:
- Tracking and detection subsystem. It finds interest object in video stream, watches and sends coordinates to gesture recognition subsystem using TCP/IP protocol.
- Gesture recognition subsystem. It receives coordinates, analyzes them and notifies user about completed gestures.

2. Analogs
Leap Motion – gesture control system – **Leap 3D**. This technology based on motion. Architecture of Leap Motion is not very complex. It has two infrared single-action and three powerful infrared light-to diode (Fig. 3).

![Figure 3. Architecture of Leap Motion](image)

Principle of work is following: infrared diodes illuminate hands, and infrared cameras make hands detection sending data to programming handler of Leap Motion. Mathematical algorithms working on program level allocate contours of hands and detect finger coordinates [6].

Another variant system is Kinect – motion controller from Microsoft. Kinect consists of two deep sensors, multicolor camera and microphone lattice (Fig. 4).

Software makes full 3d-dimension scanning motions of body, facial expressions and voice. Deep sensor consists of infrared projector with monochromatic matrix that allows to get 3d-dimension picture in different lightings.

![Figure 4. Architecture of Kinect](image)

Computing the human position is two-part process. Firstly, it calculates deep map with help infrared emitter and receiver (deep sensors). After this operation Kinect computes human position (with machine learning algorithms) [7].

Can be useful another variant neural network model of artificial intelligence for handwriting recognition. This neural networks are used for example for handwriting recognition [8] and forensic handwriting examination [9].

3. System Architecture
System consists from several related subsystems. We consider these components separately and together.
3.1. Detection system
We decide to use convolutional neural networks to solve the problem of pattern recognition, since at the moment they are superior in all parameters to other methods of solving this problem. But before you recognize a specific image, you need to find its location on the source image and translate it into a data format that is fed to the input of the convolutional network. With this in mind, the final algorithm looks like this:

1. Find the location of images for recognition on the source image
2. Convert found images to the input format of a convolutional network
3. Run the image recognition and wait for the result

We use the Connected-component labeling algorithm in the version with two full passes through the image to find image locations for recognition. The two-pass version is based on the SNM data structure (the System of Disjoint Sets).

Connected-component labeling works only on a monochrome image, so before you use it, the binarization task of the image arises. Given that the program will work on mobile devices, it was decided to use adaptive binarization, since often the lighting conditions were different in the source image. Before applying binarization, you need to rid the image of random noise. This goal is performed by a Gaussian Filter with a 3-core kernel, which is applied to the source image before binarization. After binarization, a Closing Operator with a core size of 3 elliptic type is also used to remove small tears in the images for recognition.

The single-channel image 28x28 is fed to the input of the convolutional network. But if you bring all the images to this size, then the recognition accuracy will drop, since the neural network was trained on images where the content does not occupy the entire area. Therefore, the found areas for recognition change the size by 24x24 with all the proportions preserved, and then frames with a width of 2 empty pixels on each side are added.

The architecture of the convolutional network was chosen taking into account the computing power of mobile devices and the maximum accuracy available on them. As a result, the following proven architecture was chosen:

1) Convolution layer: kernel size: 5x5, input channels: 1, output channels: 32
2) The maximum pooling layer: kernel size: 2x2, step 2x2
3) Convolution layer: kernel size: 5x5, input channels: 32, output channels: 64
4) The maximum pooling layer: kernel size: 2x2, step 2x2
5) Fully-knit layer: input channels: 64, output channels: 1024
6) Fully-knit layer: input channels: 1024, output channels: 10
7) Softmax filter

![Figure 5. The architecture of the convolutional network](image)

3.2 Gesture recognition system
At the first version this subsystem data was analyzed via analytical methods.
Later, we decided that the best solution is using neural nets for analyzing trajectories of object.

3.2.1 Analytic algorithms of first system version
We determine gesture such as set of regions that must be visited object for gesture accomplishment.

In the simplest case this regions are circles with particular coordinates and radius in 2-dimension area (in 3D it will be sphere).

Gesture - ccontinuous action, so we can’t to determine what kind of gesture was performed basing only on current object position. We must store passed object positions and check this array.

Classical algorithm is a simplest algorithm for gesture detection. Object must visit all regions of gesture in particular order.

Algorithm is very simple for realization but it very much depends from coordinates of regions. We can solve this problem with normalization of trajectory and gesture in library of gestures.

Scanning algorithm - algorithm for gesture detection based on comparing normalizing trajectories with etalon sequence of positions.

Let, solving task is input command to computer with dynamic gestures of hand. We choose 12 different action of hand as etalons (Fig. 6).

![Figure 6. Selected etalon gestures](image)

In case dynamic gestures the task of dentification becomes more complicated in view of indefinite start and end of gesture.

For solving this problem we can analyze multiple trajectories duration from 0.7 to 2.0 seconds. When system detects probable start and end of gesture we will make next operation:

1. Creating etalon gestures
2. Comparing trajectory with etalons.

Comparing gestures we can make via comparing OX, OY trajectories (Fig. 7)

![Figure 7. (a) Gesture trajectories. (b) OX trajectory (c) OY trajectory.](image)

It is necessary to normalize the unfolding of the trajectory of the produced gesture and then compare with the standards using the threshold difference-quadratic function (if applicable \(\sum(y_1 - y_2)^2 < M\sum(y_1 - y_2)^2 < M\), the gesture can be considered completed) or by using the DTW-technique (date time warping) of comparing [17] time series of different lengths.

3.2.2 The second neural net
Analitical algorithms is simple for realization but they require many resources (especially DTW-algorithm). For analysis trajectory one of the good methods is neural nets algorithms, because they is oriented for classification images.

Trajectory of object in video stream we may present as bit picture or bit matrix. We mark pixels is “1” where has been object and “0” where object wasn’t. Example graphic image of bit matrix is presented on Fig. 8.
After that we transfer bit matrix to neural net. Previously we must teach neural net using etalons with noise. Neural net analyze input images and notify about completed gestures.

Architecture of neural net is very simple. Simple one-layer perceptron provides logic of detection (Fig.9).

![Figure 8. Example of bit matrix](image)

Each neuron of this net present own gesture. If several different neurons will shoot we must choose the neuron with max activation energy (max value).

Training set is presented by etalon gestures with noise.

4. Conclusion
As a result of the work, we developed the system architecture, which has a number of positive qualities, such as cross-platform, low resource consumption and help in improving the current version of the system running on cascading algorithms (TLD-tracking).

Further improvement will increase the efficiency index and expand the possibilities of the system application in the future.

After performing the necessary improvements, we will perform detailed testing of the subsystems (checking the speed and quality of work on various hardware platforms, a comparative analysis of the effectiveness of recognition of dynamic gestures.

References
[1] Dumin A A, Sorocoumov P S, Urvanov G A and Chepin E V 2013 Architecture and prototype the system of BCI with mobile robototechnic device Messenger of national research university MEPHI
[2] Alomari M H, Samaha A and AlKamha K 2013 International Journal of Advanced Computer Science and Applications 2 pp 207-212
[3] Dumin A A, Urvanov G A and Chepin E V 2015 System of relationship with human as agent of mobile robototechnic complex. Program system and computing methods NB-Media
[4] Ren Z, Meng J, Yuan J and Zhang Z Autor of translate: A.S. Penkov Link: Robust Hand Gesture Recognition with Kinect Sensor. Web-site: http://www.uran.donetsk.ua/~masters/2013/fknt/penkov/library/article10.htm (date of handling 12.10.2015)
[5] Vasilyev A “Overview: How to arranged the technology of gesture recognition” from 30.03.2014. Web-site: http://www.mobiledevice.ru/69902-raspoznavaniia-zhestov-Technology-realsense-cvs-allsee-kin.asp (date of handling 12.10.2015)

[6] Motions controller “Leapmotion”. Web-site: https://www.leapmotion.com/(date of handling 10.10.2015).

[7] Motions controller “Kinect”. Web-site: https://msdn.microsoft.com/en-us/library/jj131033(date of handling 12.10.2015).

[8] Kulik S D 2015 Journal of Theoretical and Applied Information Technology 73 2 pp 202-211

[9] Kulik S and Nikonets D 2016 The Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (Moscow Russia) pp 221-226 Doi: 10.1109/DIPDMWC.2016.7529393

[10] Program system “Flutter”. Web-site: https://flutterapp.com/(date of handling 12.10.2015).

[11] Программа “NPointer”. Web-site: http://www.neurotechnology.com/npointer.html(date of handling 12.10.2015)

[12] Hardware and software complex «AirGesture» Web-site: http://www.theverge.com/2012/10/3/3447800/pioneer-gesture-control-car-navigation (date of handling 12.10.2015)

[13] System of gestures recognition from company “Atanor”. Web-site: http://www.atanor.ru/engineering/systema-raspoznavania-zhestov/ (date of handling 12.10.2015)

[14] Gamebox “nTobeBox”. Web-site: http://www.innodigital.net (date of handling 12.10.2015)

[15] System of gestures recognition from company “NEC”. Web-site: https://www.nec-display-solutions.com/p/ru/ru/company/about.xhtml (date of handling 12.10.2015)

[16] STB (Set-Top-Box) Web-site: http://mediapure.ru/online-servisy/chto-takoe-stb-set-top-box-iptv-

[17] http://jmlida.org/papers/doc/2011/no1/Romanenko2011Dynamic.pdf -DTW algorithm (date of handling 12.10.2015)