Using virtual data for training deep model for hand gesture recognition

E I Nikolaev, P V Dvoryaninov, Y Y Lensky and N S Drozdovsky
North-Caucasus Federal University, 2, Kulakov Ave., Stavropol, 355029, Russia

E-mail: notdeveloper@gmail.com

Abstract. Deep learning has shown real promise for the classification efficiency for hand gesture recognition problems. In this paper, the authors present experimental results for a deeply-trained model for hand gesture recognition through the use of hand images. The authors have trained two deep convolutional neural networks. The first architecture produces the hand position as a 2D-vector by input hand image. The second one predicts the hand gesture class for the input image. The first proposed architecture produces state of the art results with an accuracy rate of 89% and the second architecture with split input produces accuracy rate of 85.2%. In this paper, the authors also propose using virtual data for training a supervised deep model. Such technique is aimed to avoid using original labelled images in the training process. The interest of this method in data preparation is motivated by the need to overcome one of the main challenges of deep supervised learning: using a copious amount of labelled data during training.

1. Introduction
There has been a growing focus on improving interaction between humans and machines by allowing this to happen in a natural manner. One way to enable this natural interaction is to allow the machine to recognize the hand gesture of the user and to generate the appropriate response. Nowadays, the study on hand gesture recognition, in general, involved three aspects: gestures, chains of the gestures, multi-gestures, and interactions. Gesture refers to a static state which is about a certain position of the human hand such as a fist or an open palm. A chain generally consists of several sequential static gestures such as showing a square by open palm or circle by the fist. Multi-gesture consists of several gestures from different cameras. Interaction indicates a correlative action involving a person and an object such as touching. The overview process of handling these tasks consists of data preparation, feature extraction, hand gesture learning, and classification.

Various intelligent techniques have been used to perform hand gesture recognition such as hidden Markov models [1-3], support vector machines [4, 5] and neural networks [6, 7]. Deep convolutional neural networks (CNN) has shown real promise for the classification efficiency for hand gesture recognition problems [8-10]. In the recent years, many researchers have found different ways to solve the problems of gesture recognition with Kinect [11, 12]. The using of depth maps and Kinect technology in this paper are avoided and proposed deep model works with a data obtained from the web camera. Moreover, the two-phase deep architecture based on CNN is proposed. This model achieves a classification of thirty gestures of the human hand.

Deep learning uses a very deep layered network and a huge number of data, so preparing and labelling data are serious problems. To avoid a lack of labelled data, virtual datasets for training are used. It is significantly important not only in gesture recognition systems but also in such areas as
human pose estimation, emotion prediction and driverless cars. In this paper, a method of generating the virtual data is proposed. It’s based on constructing a 3D-model of human hand and immersion it in a virtual environment with additional information about data labels.

2. Overview of the hand gesture recognition system
The whole architecture of the system for gesture recognition is described in this section. The presented system (Figure 1) consists of the virtual dataset generator (DSG), the feature extraction module (FEM), the classification module (CM) and the model for high-level interpretation (HLM).

![Figure 1. An architecture of the system for hand gesture recognition.](image)

The feature extractor module is presented by the sequence of convolutional and pooling layers. The authors use a vgg16 model [13] without full connection layers. The classification problems are solved by two separate CMs. The first CM (FC1) consists of three fully-connected layers and it evaluates a three-component vector. This vector presents a hand orientation in 2D-space (the third coordinate is reserved for the future research). The second classification module (FC2) evaluates an output vector for prediction the class of hand gesture. Both CMs use the same feature extraction module.

In this paper, the real data for training purpose are not used. Instead, the using virtual datasets (training DS and testing DS) is proposed to solve one of the most significant challenges in deep learning. Therefore, in the framework of the research, it is important to develop the generator of datasets which is able to generate a huge number of training images with appropriate features.

3. Training the control system for hand gesture recognition

3.1. Datasets generation and labelling
The structure of the module for virtual data generation is shown in figure 2. The most complex stage of data preparation is a building 3D-models of the human hand.

![Figure 2. The structure of the virtual data generator.](image)

Three different 3D-models of the human hand in Blender have been created by using sculpture tool for the realistic view. In order to be able to control each finger position in automatic mode (application
for creating images realizes this stage), the virtual bones have been added in each 3D-model. The position of each finger can be changed by moving key points. In this way, it does not need to construct a 3D-model from scratch in order to obtain the image of new gesture: it is realised by adding a set of key points in the storage of key points. In order to provide a high level of generalization for CNN, the different backgrounds are used in 3D-scenes, as well different positions of camera view (that transformations were realized with OpenGL library). In fact, each item of the training dataset is a union of three elements: the coordinates of the key points; the background image, the coordinates of the camera position. In addition, the augmentation procedure is implemented during the training process.

3.2. Training deep model

Training process consists of two phases. The first phase is training CNN for hand orientation prediction, and the second one – training CNN for gesture class prediction. Training is performed on a single GeForce Titan graphics card. The main limit here is the device’s memory capacity as described training and testing datasets exceed a memory capability. The proposed model is evaluated on the dataset consisting of 630,000 data samples obtained from 3 different 3D-models of the human hand. For each model 30 different hand gestures have been created. Each of the gestures is represented by 7,000 snapshots summing up to 210,000 data samples per 3D-model. This would result in the transformation and storage of 630,000 data samples by the Tensorflow library during training for 100-on-1 cross-validation (based on samples not on models), including weights as well as the subsequent image transformation steps, which is more than the device can store during the training phase.

The number of data samples during training the deep model for hand orientation prediction is reduced to 2,000 samples per gesture, each randomly taken from the whole sample set. This still yields a training set of 180,000 hand poses – more than enough to validate the proposed approach. During training, the method of data augmentation presented in [14] has been implemented.

The resulting activations from the first convolution layer are depicted in Figure 3. The first 14 activation maps for one class (CLASS_PALM) of gesture are shown. Presented activation maps are obtained from the first pooling layer of the feature extraction module.

![Figure 3. The activation set for the first pooling layer.](image)

After training for 5,500 epochs the first model achieved its best performance producing an accuracy rate of 95.93% on the testing set. Further training with the same parameters seems to cause overfitting.

4. Experimental evaluation and results

The problem of hand segmentation is not considered in the framework of this research. In order to obtain an image of the hand from the camera in a real-time mode, the bounding rectangle is used (Figure 4).
Figure 4. Real time hand gesture recognition experiment.

Experiments were performed with web camera (with resolution 1024x768) and GeForce 750M graphics card (a low-performance laptop has been used as deployment platform). Python and Tensorflow have been used for prediction the hand signs. The output video frames are shown in Figure 4.

The evaluation of the proposed gesture recognition system has been implemented with two evaluation datasets: namely self-built virtual evaluation dataset (VED) and a self-built real evaluation dataset (RED). Dataset RED contains images of the human hand signs from the web camera. The VED was created by the same method as training and testing datasets. Table 1 illustrates parameters and the evaluation mean accuracy for both data sets.

| Images per class | Number of classes | Total number of images | Mean accuracy |
|------------------|-------------------|------------------------|---------------|
| VED 1000         | 30                | 30,000                 | 92.3%         |
| RED 100          | 10                | 1,000                  | 85.2%         |

The gap between accuracy evaluated on different validation sets is caused by different nature of real and virtual data. In spite of this, the proposed deep model demonstrates the high level of generalization ability. The comparison of the presented state-of-the-art results and the most famous investigations based on public data sets is a future work.

5. Conclusion and outlook

This paper implements the experiments of the hand gesture recognition based on the deep model with two classification modules and a shared feature extractor. The proposed hand gesture recognition system is able to predict 30 different hand poses in real-time mode. Such model can be used in different areas where the need for human-computer interaction or remote control by gestures is most urgent and acute. Experimental results show that the proposed deep model demonstrates a mean accuracy rate of 92.3% on a virtual validation data set and 85.2% on a real validation data set.

Future work will focus on the development of the event library for the hand gesture recognition. Such system will perform not just a classification task but regression: a virtual data generator can produce datasets for the training of the regression model. Additionally, future work will explore the transferability of this approach to problems of object recognition in such areas as a human pose estimation, driverless cars, and mobile robotics in order to prove its general applicability. The proposed results can easily be extended to other similar tasks allowing for improved performance of deep models under difficult circumstances, especially in the tasks with the lack of labelled data.

References

[1] Yang Z, Li Y, Chen W, Zheng Y 2012 Dynamic hand gesture recognition using Hidden Markov models Proc. of the 7th Int. Conf. on Computer Science & Education (ICCSE) (Melbourne, Australia: IEEE) pp 360-365
[2] Marcel S, Bernier O, Viallet J E and Collobert D 2000 Hand gesture recognition using input-output hidden Markov models Proc. Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition (Grenoble: IEEE) p 456
[3] Liu N, Lovell B C, Kootsookos P J and Davis R I 2004 Model structure selection & training algorithms for an HMM gesture recognition system Proc. of the 9th Int. Workshop on Frontiers in Handwriting Recognition (IFWHR-9’04) (IEEE) pp 100-105
[4] Ren Y, Zhang F 2009 Hand Gesture Recognition Based on MEB-SVM Int. Conf. on Embedded Software and Systems (IEEE) pp 344-349
[5] Nguyen T-N, Vo D-H, Huynh H-H, Meunier J 2014 Geometry-based static hand gesture recognition using support vector machine 13th Int. Conf. on Control Automation Robotics & Vision (ICARCV) (IEEE) pp 769-774
[6] Tang A, Lu K, Wang Y, Huang J and Li H 2015 A real-time hand posture recognition system using deep neural networks ACM Transactions on Intelligent Syst. and Technol. (TIST) 6(2) 21
[7] Wu D and Shao L 2015 Deep dynamic neural networks for gesture segmentation and recognition ECCV 2014 Workshops LNCS ed L Agapito M M Bronstein et al (Heidelberg: Springer) 8925 552-571
[8] Nagi J, Ducatelle F, Di Caro J A, Cireşan D, Meier U, Giusti A, Nagi F, Schmidhuber J and Gambardella L M 2011 Max-pooling convolutional neural networks for vision-based hand gesture recognition IEEE Int. Conf. on Signal and Image Processing App. (ICSIPA) (Kuala Lumpur, Malaysia: IEEE) pp 342-347
[9] Simonyan K and Zisserman A 2014 Two-stream convolutional networks for action recognition in videos Advances in Neural Information Processing Systems 27 568-576
[10] Molchanop P, Gupta S, Kim K, Kautz J 2015 Hand gesture recognition with 3D convolutional neural networks IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 1-7
[11] Maturana D, Scherer S 2015 VoxNet: a 3D convolutional neural network for real-time object recognition IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS) (IEEE) pp 922–928
[12] Li Y 2012 Hand gesture recognition using Kinect Proc. of 2012 IEEE 3rd Int. Conf. on Software Engineering and Service Science (IEEE) pp 196-199
[13] Simonyan K, Zisserman A 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition Preprint arXiv:1409.1556v6 [cs.CV]
[14] Nikolaev E I, Dvoryaninov P V, Lensky Y Y and Drozdovsky N S 2017 An Intelligent System for Content Generation Proc. of the 2nd Young Scientist’s Int. Workshop on Trends in Information Processing (YSIP2’2017), Dombar, Russian Federation ed S Hölldobler, A Malikov et al (CEUR-WS) 1837 152-157