ICIC Express Letters
Part B: Applications
Volume **8**, Number **8**, August **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 1211–1217

SEMANTIC RECOGNITION OF SIGNED LANGUAGE USING CONVOLUTIONAL NEURAL NETWORK

Lanzhong Wang

School of Foreign Languages and Literature
Shandong University
No. 27, Shanda Nanlu, Jinan 250100, P. R. China
lzwangsdu@sina.com

Received February 2017; accepted May 2017

Abstract. *In this paper we study the hand gesture recognition problem for signed language. The complex background in real world application is a major challenge for hand region segmentation. First, we adopt the skin color model to pre-process the input image. Second, skin color filtered images are used to train a fast convolutional neural network for hand region detection. The detection is converted into a regression problem which leads to more efficient detection results. Third, the detected hand image patch is applied with a state-of-the-art landmark localization algorithm using Markov Random Fields and Active Shape Models. Finally, the American signed language is recognized in this paper for verifying the proposed system. Experimental results show that compared with several other approaches, the proposed gesture recognition system achieves satisfactory performance.*

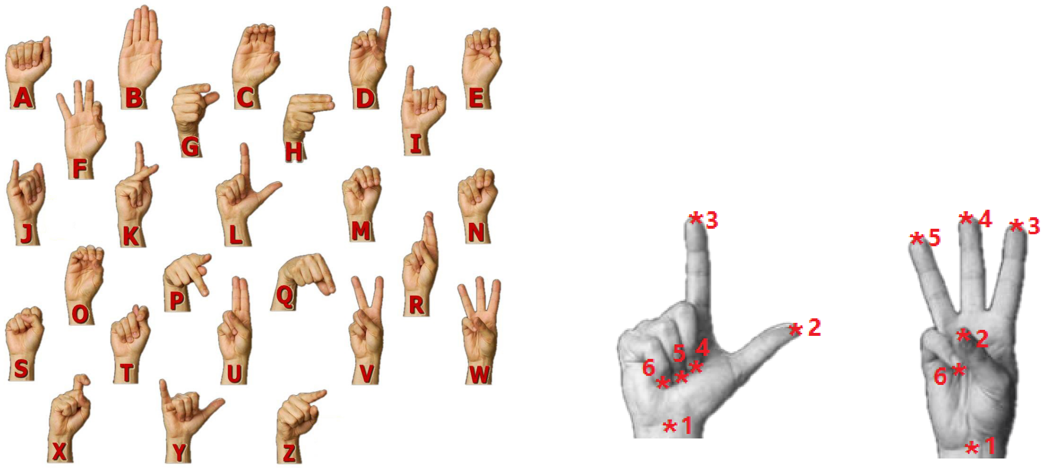
Keywords: Gesture recognition, Convolutional neural network, Signed language, Landmark localization

1. **Introduction.** Automatic recognition of signed language may help people to better interact with computers, especially for those who need hand gestures to express themselves. American signed language is shown in Figure 1(a), recognizing the semantic meaning in these gestures may innovate new applications in natural human-computer interaction.

Convolutional Neural Network (CNN) has drawn much attention in computer vision community. Many new deep structures have been proposed [1, 2]. It is an effective model for learning patterns in local regions. The layer-wise process provides good high-level abstraction. However, their computational efficiency is not satisfactory. In gesture recognition the major drawback in computational efficiency is the detection of the bounding box. Most of the current works require a sliding window to repeatedly calculate numerous local image regions.

Landmark localization is a fundamental problem in computer vision [3, 4, 5]. It is key to gesture recognition [6, 7, 8]. In this paper, the landmark points are defined on a hand, as shown in Figure 1(b). Various algorithms have been proposed in the past. Zhang et al. [9], proposed to use deep neural network to detect landmarks. Although the robustness was improved, a large amount of data was also required and the computational efficiency was not satisfactory. Zhou et al. [10], used the convolutional network in a coarse-to-fine framework and it was robust against different backgrounds. However, the computational burden was relatively high. Sangineto [11] proposed to use the dense-SURF feature to locate landmarks. The accuracy was improved compared to traditional methods, while the occlusion was still a challenging problem.

In this paper we propose to use a novel convolutional neural network to solve the hand ROI (Region of Interest) detection problem as a regression problem. Pixel level information is used directly to estimate the target. The robust landmark localization is then used to improve the gesture recognition performance. The contributions of this paper



(a) (b)

Figure 1. Examples of hand gestures: (a) American signed language; (b) defined landmark points on hand gestures

are two folds: i) the proposed regression CNN approach is used to detect objects without a bounding box with very low computational burden; ii) the landmark localization accuracy is significantly improved due to the good initialization using the proposed regression CNN approach. The rest of the paper is organized as follows: Section 2 describes the hand ROI detection method; Section 3 describes the robust landmark detection algorithms; in Section 4, the gesture recognition model is introduced; in Section 5, experimental results are provided; finally, the conclusion is drawn in Section 6.

2. Hand ROI Detection Using Regression-CNN.

The hand ROI is detected by two steps. First, the skin color is modelled and the pixels are labelled when the color is close to human skin. Second, the hand ROI is modelled by deep convolutional neural network as a regression problem (Regression-CNN).

The distance is measured by Euclid distance in the RGB (Red, Green, Blue) color space and a threshold th^d_{skin} is used to decide whether the pixel belongs to human skin. A further skin classification is carried out for improved accuracy. The HSV (Hue, Saturation, Value) color space, YCbCr color space and TSL (Tint, Saturation, Lightness) color space are used.

TSL color space is defined as the following [12].

$$T = \begin{cases} \dfrac{1}{2\pi}\arctan\dfrac{r'}{g'} + \dfrac{1}{4} & g' > 0 \\ \dfrac{1}{2\pi}\arctan\dfrac{r'}{g'} + \dfrac{3}{4} & g' < 0 \\ 0 & g' = 0 \end{cases} \tag{1}$$

$$S = \sqrt{\frac{9}{5}(r'^2 + g'^2)} \tag{2}$$

$$L = 0.299R + 0.587G + 0.114B \tag{3}$$

where $r' = r - 1/3$, $g' = g - 1/3$, $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$.

The thresholds that are used for the fine skin classification are achieved from a large testing dataset. The rules are manually adjusted for the acceptance of skin color as the following:

* Rule No.1: $Y > 60$ & $80 < Cb < 135$ & $130 < Cr < 175$ in YCbCr space.

* Rule No.2: $0 < H < 410$ & $0.14 < S < 0.9$ in HSV space.
* Rule No.3: $0.4 < T < 0.8$ & $0.02 < S < 0.5$ & $L > 60$ in TSL space.

The hand ROI detection is the bottle neck of our gesture recognition system due to the complex backgrounds interference and the unstable performance of the traditional skin color detection methods. The hand detection based on the skin color is used as an initial step to pre-process the raw image. The classifier based on the CNN model is able to detect the hand ROI based on a relatively large number of training samples.

In this section, we propose a novel fast convolution neural network that directly detects the ROI without traditional sliding window. In traditional hand detection methods a sliding window is needed to cut out the local image patch and then apply the classification algorithm. This process will consume significant computing resources. In our approach, the image is sent into the neural network as a whole. The ROI is represented by four values, two center coordinates, the width and the height. The tensor is represented as: $\boldsymbol{f} = \{x, y, w, h\}$.

The neural network is then used to solve this regression problem directly based on the pixel values.

The initial convolutional layers are designed to process the raw input data and extract meaningful features. The fully connected layers do the regression work and predict the tensor.

The neural network architecture consists of 6 convolutional layers and 2 fully connected layers. The activation mechanism follows the leaky ReLU (rectified linear units):

$$f(x) = \begin{cases} x, & x > 0 \\ 0.1x, & \text{otherwise} \end{cases} \tag{4}$$

It allows a small, and non-zero gradient when the neuron is not activated as shown in Figure 2 compared to standard ReLU and softplus approximation.

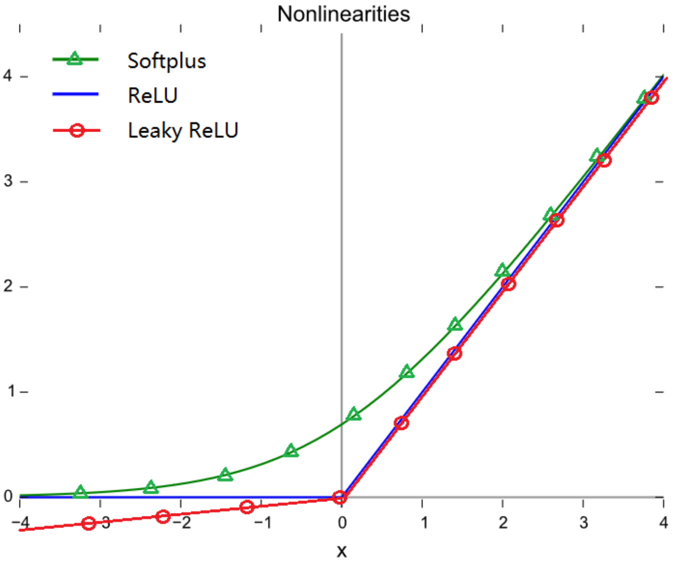


Figure 2. Standard rectified linear units, leaky ReLU and softplus

During the training of the proposed model architecture, we optimize the following target function:

$$J = \sum_{i=0}^{M} I_i \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] + w \sum_{i=0}^{M} I_i \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \tag{5}$$

where M is the number of samples, w is the weight parameter of the second part penalty, and I_i indicates the appearance of the hand target in ith image.

3. **Landmark Detection.** Our landmark detection module is built upon the Active Shape Model (ASM) [13]. The drawback of ASM is that it requires a good initial position to start searching. In many hand gesture recognition problems the initial position is very difficult to achieve. The complex backgrounds and illumination changes make this task more challenging.

Regression-CNN is used to initialize the ASM. The regression results are used to calculate the prior of landmark locations. High Order Markov Random Fields (MRF) [14] is used to configure the landmark detection results based on a pre-learnt shape model. For occluded landmarks, MRF can be used to predict the missing landmark coordinates.

The ASM is trained on an image set with annotated landmarks. These landmarks provide the ground-truth and they are often manually annotated. The algorithm consists of two sub-models, one is the profile model, and the other is the shape model. In the landmark detection process, each landmark is first located independently and then the locations are corrected according to the shape model in an iterative fashion.

The shape model is represented as:

$$\hat{\boldsymbol{x}} = \overline{\boldsymbol{x}} + \boldsymbol{\Phi b} \tag{6}$$

where $\hat{\boldsymbol{x}}$ is the shape vector generated from hand landmarks, $\overline{\boldsymbol{x}}$ is the mean shape vector, $\boldsymbol{\Phi}$ contains the eigenvectors of covariance matrix from the distortion and $\boldsymbol{b}$ is a parameter for generating different shapes.

4. **Gesture Recognition.** The landmark coordinate is directly correlated with the finger positions. The rotation changes may influence the landmark coordinates and bring a large variance in data samples. In some occlusion cases the coordinates may be missing or mistaken.

In the gesture recognition model we adopt the traditional Support Vector Machine (SVM) [16] to learn the difference between each gesture. Since the training samples are much smaller compared to the hand ROI detection or landmark detection problem, the SVM solution achieves the best results compared with neural networks.

The feature vector constructed for recognition contains two parts, the landmark coordinates and the Local Binary Pattern (LBP) features. Auto-encoder [15] is used to reduce the dimensionality.

An auto-encoder is a deep neural network structure that combines several layers of Restricted Boltzmann Machines (RBMs). The input features are encoded throughout the network and the dimensionality is forced to be reduced during the process. The reduced features are represented in neurons and decoded through the similar network structure to recover the original features. The optimization is taken place during the training stage to ensure that the recovered features are as close as possible to the input features.

The key idea to tune our SVM classifier is preventing the error propagation in the multi-class SVM decision tree. A total number of C_{26}^2 two-class classifiers are investigated, and the error matrix between any two gestures is achieved as shown in Table 1.

Based on the two-class classification results we can design a decision tree that follows the "one-against-all" approach, in which the more reliable classification is placed in the earlier stage of the decision tree. In other words, the most difficult one is recognized at last in order to prevent the error propagation in the decision tree. The order of the gestures in the "one-against-all" tree is ranked according to the error rate in the error matrix.

5. **Experimental Result.** We carry out three experiments to verify the effectiveness of the proposed gesture recognition system. First, we test the localization accuracy of hand ROI (Region of Interest) detection. Second, we test the accuracy of our landmark detection algorithm using Regression-CNN initialization. Third, the recognition rates of signed language gesture using various algorithms are compared.

Table 1. Error matrix of the two-class classification rate (%) (the highest 10 rates are listed.)

Pairs	B	E	V	L	W	D	Y	Z
A	0.2	0.8	0.9	1.1	1.3	1.5	1.7	1.9
W	1.6	1.5	1.7	1.7	1.8	1.6	2.1	2.1
H	1.9	1.9	1.8	2.1	2.1	2.1	2.3	1.9
N	1.9	2.0	2.2	1.8	2.3	2.3	2.4	2.4
B	2.1	2.2	2.3	2.4	1.8	2.3	2.1	1.9
Z	2.2	2.3	2.3	2.8	2.1	2.0	2.2	2.1
T	2.5	2.6	2.2	2.3	2.1	2.3	2.0	2.4
F	2.3	2.2	2.2	2.3	2.3	2.5	2.2	2.1

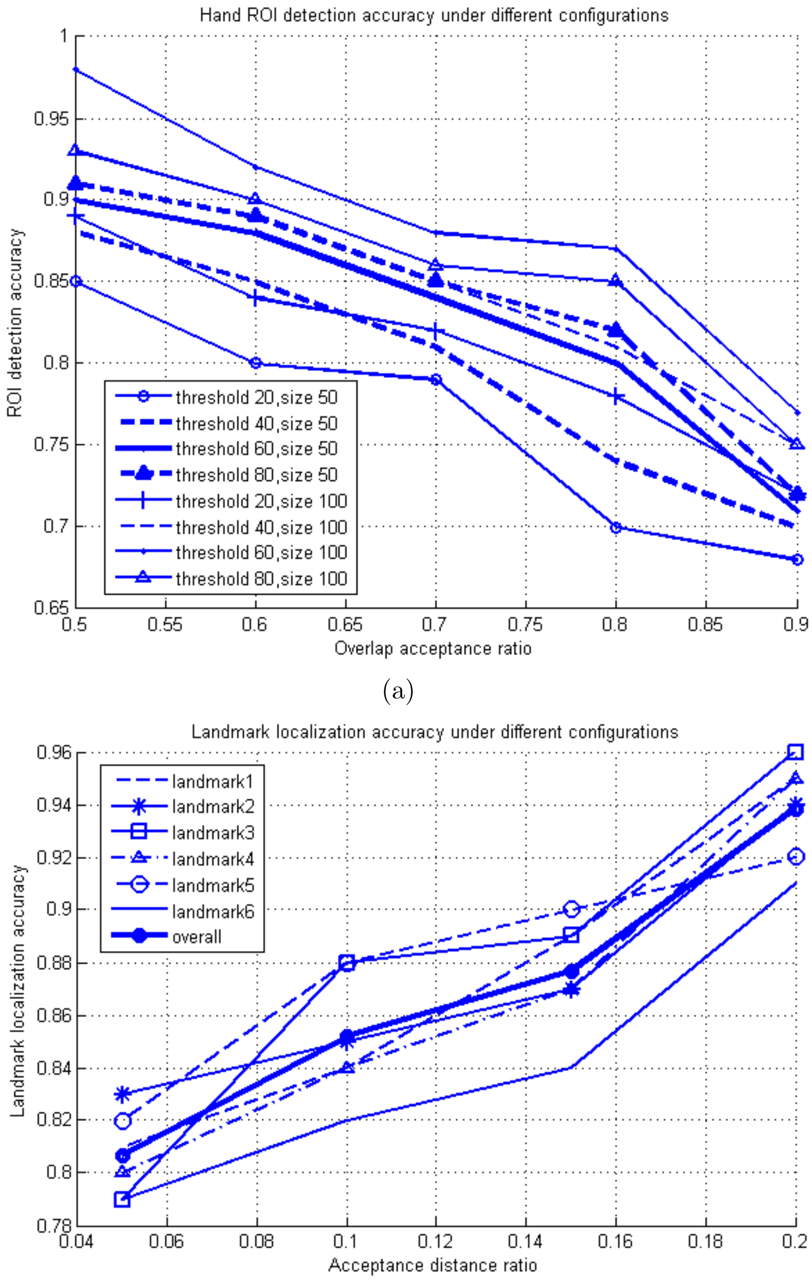
As shown in Figure 3(a), the hand ROI detection rate is dependent on the threshold th^d_{skin} in the skin color detection stage and the training set size s^t in Regression-CNN model. The accuracy is measured by the overlap area of the hand ROI. A success detection of the ROI is defined by the ratio of overlap area. Generally above 80% overlap between the detected ROI and the ground-truth ROI is accepted in practical applications. From Figure 3(a), we can see that the success rate changes along the overlap acceptance ratio.

The landmark points are defined on a hand, as shown in Figure 1(b). The occlusion is a major challenge in our cases. The global geometric constraints are used for modelling the relations between landmarks. Prediction based on the proposed landmark detection algorithm may help the accurate location of finger tips.

As shown in Figure 3(b), the landmark localization accuracy is demonstrated. The Y-axis is the percentage of successfully located landmark and the X-axis is the acceptance distance ratio. The distance ratio is defined as $|\boldsymbol{P_l} - \boldsymbol{P_g}|_{L2}/(W + H) * 2$, where $\boldsymbol{P_l}$ and $\boldsymbol{P_g}$ are the detected landmark location and the ground-truth landmark location, W and H are the width and height of hand ROI respectively. The acceptance distance ratio is a threshold and any landmark localization error smaller than this ratio is considered as a successful detection. We can see that the proposed Regression-CNN initialization for ASM model is successful. This proves the effectiveness of our proposed landmark detection algorithm.

The final signed language recognition result is shown in Table 2. Notice that Method-I uses human annotation and it is not fully automatic. The proposed system is compared with a number of other approaches. In Method-I, we use the ground-truth ROI provided by the human annotator instead of an automatic detection. In Method-II, only the traditional skin color based model is used to find the hand and the Regression-CNN model is not used for assistance. In Method-III, the landmark coordinates are not combined with texture features for SVM classification. Model-I to Model-III are described in Sections 2-4, and they are combinations of our proposed modules. In Method-IV, the Region-Proposal-CNN [17] model is adopted for comparison instead of the proposed Regression-CNN model in our paper.

We can see that by removing the hand ROI detection, we need a significant computing time due to the bad initialization in the landmark localization. The accuracy of the proposed algorithm is the closest one to Method-II in which the human annotated ROIs. The speed of the proposed Regression-CNN is 310ms per image while the speed of the Method-IV which uses the state-of-the-art fast R-CNN needs 810ms to process one image under the same testing condition.



(a)

(b)

Figure 3. Detection accuracy tests, (a) hand ROI detection accuracy; (b) landmark localization accuracy

Table 2. Final recognition accuracy (%) on signed languages

Methods	False Acceptance Rate	False Rejection Rate	Averaged Error Rate	Time
Proposed System	9.1	8.3	8.7	310ms
Method-I	8.9	8.1	8.5	110ms
Method-II	13.2	17.2	15.2	1010ms
Method-III	16.3	15.1	15.7	290ms
Method-IV	9.3	8.5	8.9	810ms

6. Conclusion.

In this paper we treat the image detection as a regression problem. Convolutional neural network is used to solve the regression problem with reliable performance on hand ROI detection. The detected hand region is then used for landmark localization based on ASM and MRF. The proposed semantic recognition system is successful compared to several traditional approaches. Since the regression is performed directly on the pixel level, the bounding box is not necessary. This improves the computation speed greatly. In future work, we will further extend the Regression-CNN to other fast object tracking applications.

REFERENCES

[1] Y. Jia, E. Shelhamer, J. Donahue et al., Caffe: Convolutional architecture for fast feature embedding, *Proc. of the 22nd ACM International Conference on Multimedia*, pp.675-678, 2014.

[2] R. Girshick, J. Donahue, T. Darrell et al., Rich feature hierarchies for accurate object detection and semantic segmentation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.580-587, 2014.

[3] M. P. Segundo, C. Queirolo, O. Bellon et al., Automatic 3D facial segmentation and landmark detection, *IEEE the 14th International Conference on Image Analysis and Processing*, pp.431-436, 2007.

[4] Y. Huang, H. Yao, S. Zhao and Y. Zhang, Towards more efficient and flexible face image deblurring using robust salient face landmark detection, *Multimedia Tools and Applications*, pp.1-20, 2015.

[5] D. Han, Y. Gao, G. Wu et al., Robust anatomical landmark detection with application to MR brain image registration, *Computerized Medical Imaging and Graphics*, vol.46, pp.277-290, 2015.

[6] Z. Ren, J. Yuan, J. Meng and Z. Zhang, Robust part-based hand gesture recognition using kinect sensor, *IEEE Trans. Multimedia*, vol.15, no.5, pp.1110-1120, 2013.

[7] S. S. Rautaray and A. Agrawal, Vision based hand gesture recognition for human computer interaction: A survey, *Artificial Intelligence Review*, vol.43, no.1, pp.1-54, 2015.

[8] Q. Pu, S. Gupta, S. Gollakota and S. Patel, Whole-home gesture recognition using wireless signals, *Proc. of the 19th Annual International Conference on Mobile Computing & Networking*, pp.27-38, 2013.

[9] Z. Zhang, P. Luo, C. C. Loy and X. Tang, Facial landmark detection by deep multi-task learning, *Proc. of European Conference on Computer Vision*, pp.94-108, 2014.

[10] E. Zhou, H. Fan, Z. Cao et al., Extensive facial landmark localization with coarse-to-fine convolutional network cascade, *Proc. of the IEEE International Conference on Computer Vision Workshops*, pp.386-391, 2013.

[11] E. Sangineto, Pose and expression independent facial landmark localization using dense-SURF and the Hausdorff distance, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.35, no.3, pp.624-638, 2013.

[12] J.-C. Terrillon and S. Akamatsu, Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments, *Proc. of the 3rd International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, pp.130-135, 1998.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, Active shape models – Their training and application, *Computer Vision and Image Understanding*, vol.61, no.1, pp.38-59, 1996.

[14] Z. Jiang and C. Huang, High-order Markov random fields and their applications in cross-language speech recognition, *Cybernetics and Information Technologies*, vol.15, no.4, pp.50-57, 2015.

[15] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, vol.313, no.5786, pp.504-507, 2006.

[16] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt and B. Scholkopf, Support vector machines, *IEEE Intelligent Systems and their Applications*, vol.13, no.4, pp.18-28, 1998.

[17] S. Ren, K. He, R. Girshick et al., Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, pp.91-99, 2015.