An effective sign language learning with object detection based ROI segmentation

Sunmok Kim, Yangho Ji, and Ki-Baek Lee
Dept. Electrical Engineering
Kwangwoon University
Seoul, South Korea
kblee@kw.ac.kr

Abstract— **This paper proposes a novel sign language learning method which employs region of interest (ROI) segmentation preprocessing of input data through an object detection network. As the input, 2D image frames are sampled and concatenated into a wide image. From the image, ROI is segmented by detecting and extracting the area of hands, crucial information in sign language. The hand area detection process is implemented with a well-known object detection network, you only look once (YOLO) and the sign language learning is implemented with a convolutional neural network (CNN). 12 sign gestures are tested through a 2D camera. The results show that, compared to the method without ROI segmentation, the accuracy is increased by 12% (from 86% to 98%) as well as the training time is reduced by over 50%. Above all, through the pretrained hand features, it has the advantage of ease in adding more sign gestures to learn.**

Keywords-sign language; video recognition; machine learning; CNN; object detection

I. INTRODUCTION

There have been many studies about sign language recognition for communicating with the hearing impaired. The most popular approach is that the trajectories of hands are extracting by using haptic devices [1-2]. Instead of the devices, RGB-D cameras for the depth information have been widely used [3-7]. However, there exist still some drawbacks in the previous research. It is inconvenient to wear the haptic devices such as gloves. RGB-D cameras have some limitations of price, resolution, and effective range. Furthermore, these two approaches have the same problem in common that the trajectory needs to be defined separately every time a gesture is added.

Therefore, recently, there have been machine learning based approaches with convolutional neural network (CNN) and low-cost 2D camera [8-11]. However, even though their innovative result, it has not been easy to extend the problem into practical domain since they require a lot of training data. The reason of inevitable massive data is that, we need to have the network know that hand gesture patterns are the key features for the classification under the limited information of the hand area whose size is relatively insignificant compared to the entire image size.

To overcome the ineffectiveness mentioned above, a novel method of learning sign language is proposed in this paper by segmenting the hand areas from an image as the region of interest (ROI). In the proposed method, hands are considered as objects and as the detection network for them, YOLO is employed, which is a well-known real-time object detection network scheme [12]. Once the hand areas are distinguished, background becomes negligible and there is no need to put massive data of various situations. In other words, the network does not have to learn which area has more valuable information. It does not matter which clothes you wear. Consequently, accuracy is enhanced, and training speed is accelerated. Also, since we do not need various situation data, it is not difficult to add data for expansion.

This paper is organized as follows. Section 2 explains the proposed methodology in detail. In Section 3, the experimental results of the proposed system are demonstrated through 12 different sign gestures. Lastly, Section 4 presents conclusions.

II. THE PROPOSED METHOD

The proposed sign language learning method is divided into two parts as shown in Fig. 1. The former part is the ROI segmentation through an object detection network and the latter part is the sign language learning through a classification network. There are two advantages of this method compared to the conventional one. First, the problem becomes more simple and clear since the hand area is segmented. Thus, the network performance and the training speed can be improved. Second, through hand area segmentation, other variables such as backgrounds and clothes can be ignored. This means that the proposed method requires relatively less data then before and it is also more convenient to learn additional sign gestures.

A. ROI segmentation

At first, a wide image is created by sampling and concatenating the original video frames. And then, by using network that detects the hand area, an ROI segmented image that contains only hand area is obtained. As the object detection network, you only look once (YOLO) is employed [14]. The YOLO has relatively faster learning speed as well as high accuracy. The detailed structure of the object detection network is shown in the upper part of Fig. 1. The concatenated image enters the input of the object detection network and comes out as $13\times13\times30$ feature maps which contains the information vectors for 13×13 cells. An information vector V is defined as

0-7695-6370-8/18/31.00 ©2018 IEEE
DOI 10.1109/IRC.2018.00069

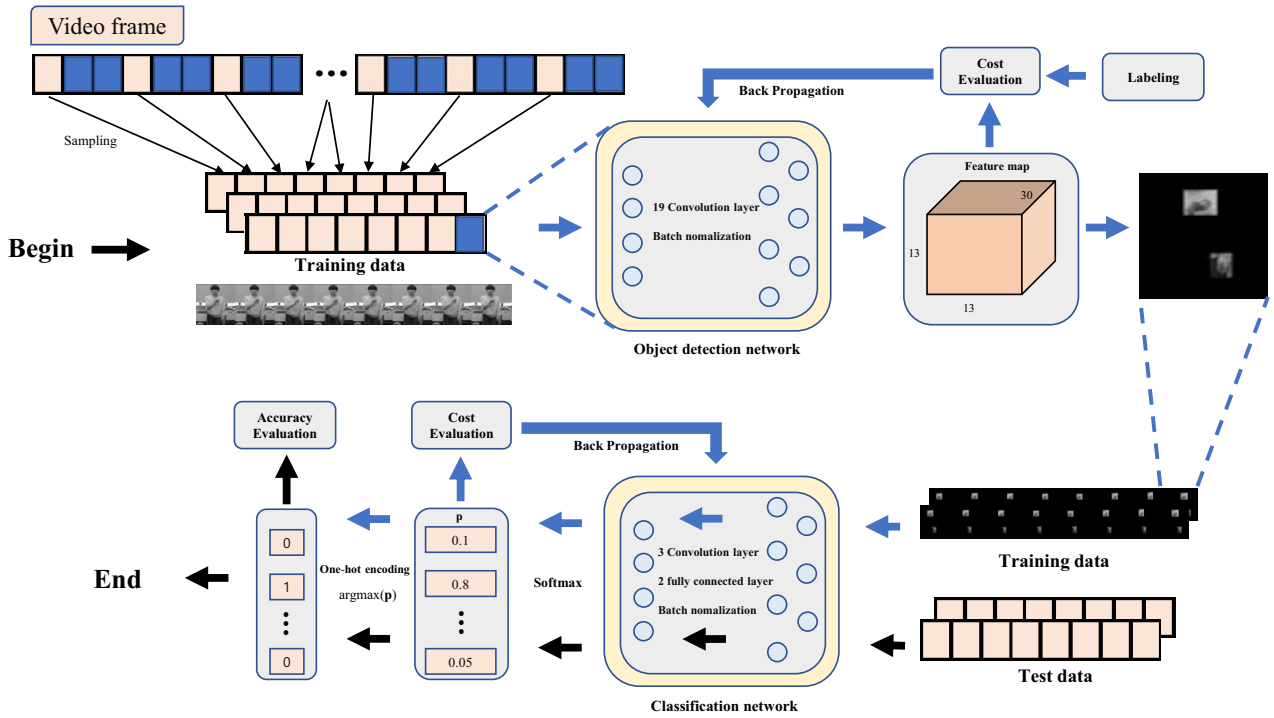

Figure 1. Flow diagram of the proposed method

following:

$$V=\{c_1, l_1, x_1, y_1, w_1, h_1, \ldots, c_K, l_K, x_K, y_K, w_K, h_K\} \quad (1)$$

where K is the number of candidate anchor boxes and c_k, l_k, x_k, y_k, w_k, and h_k are confidence level, class, x position, y position, width, and height of the k-th anchor box, respectively. Since we set K as 5, the total length of V is 30 as mentioned above. Confidence level is the probability of an object in the box, which is for removing overlapped boxes. By using the feature maps, the ROI segmented image can be made by turning the areas except the hand areas into black. To improve the training speed and performance, the weight of the YOLO network is set as the initial weight of the pretrained network with the PASCAL VOC 2007 data set.

B. *Sign language learning*

In the sign language learning process, the ROI segmented image enters input of the classification network. As the classification network, convolutional neural network is used, which is known to perform well for image classification problems [13-15]. The network consists of three convolution layers and two fully-connected networks. The goal of the classification network is obtaining the probability vectors of the classes. The overall process is shown in the lower part of Fig. 1. As can be seen in Fig. 1, since the input image only consists of hands, the goal of the learning becomes clear. First, input data passes the network and comes out as and the probability vectors. By using this, the cost is calculated as following:

$$C = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log p_n + (1-y_n)\log(1-p_n)\right] \quad (2)$$

where N is the length of one-hot encoded label vector, y_n is the n-th element of the label vector, and p_n is the n-th element of the probability vector $\boldsymbol{p}$. After that, the weights of the network are updated through back propagation method for each iteration. Learning is completed after a specified number of iterations and the accuracy can be evaluated using the test data. And for the training speed and stability, we used batch normalization technique and Xavier initialization method [16-18].

III. EXPERIMENTS

This section explains about the data sets used in the experiment. The result of the experiment and its analysis are followed.

A. *Data set*

The training set consists of the images for 12 gestures in 60 situations taken from 1.0-m distance. Situations mean the combinations of backgrounds, clothes, etc. The test set includes the images taken from two different distances, 1.0 m (Fig. 2) and 1.5 m (Fig. 3) for the assumption of tough test environment. Labels from 1 to 12 are assigned to the gestures and they are encoded into one-hot vectors as shown in the Table 1.

B. *Result*

Fig. 4 and Fig. 5 show the cost and accuracy over batch number with the 1.0-m and 1.5-m distance test sets, respectively.

Figure 2. 1.0-m distance test set image

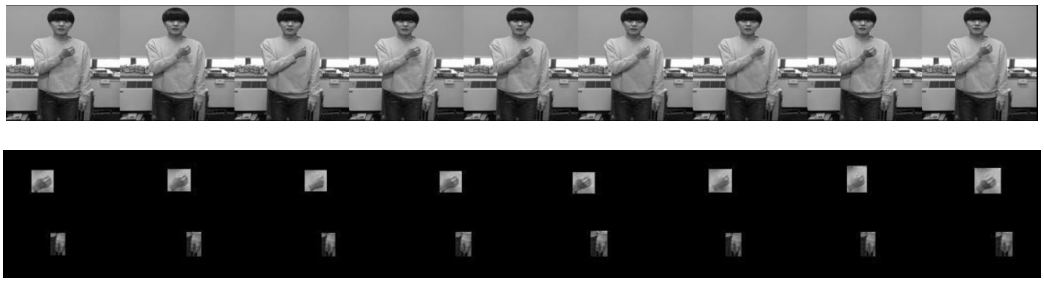

Figure 3. 1.5-m distance test set image

TABLE I. DATA LABELING

Meaning	Lable	One-hot vector
Love	1	(1,0,0,0,0,0,0,0,0,0,0)
Thank	2	(0,1,0,0,0,0,0,0,0,0,0)
Happy	3	(0,0,1,0,0,0,0,0,0,0,0)
Effort	4	(0,0,0,1,0,0,0,0,0,0,0)
Regrettable	5	(0,0,0,0,1,0,0,0,0,0,0)
Give	6	(0,0,0,0,0,1,0,0,0,0,0)
Apologize	7	(0,0,0,0,0,0,1,0,0,0,0)
Stuffy	8	(0,0,0,0,0,0,0,1,0,0,0)
Same	9	(0,0,0,0,0,0,0,0,1,0,0)
Learn	10	(0,0,0,0,0,0,0,0,0,1,0)
Move	11	(0,0,0,0,0,0,0,0,0,0,1)
Funny	12	(0,0,0,0,0,0,0,0,0,0,0)

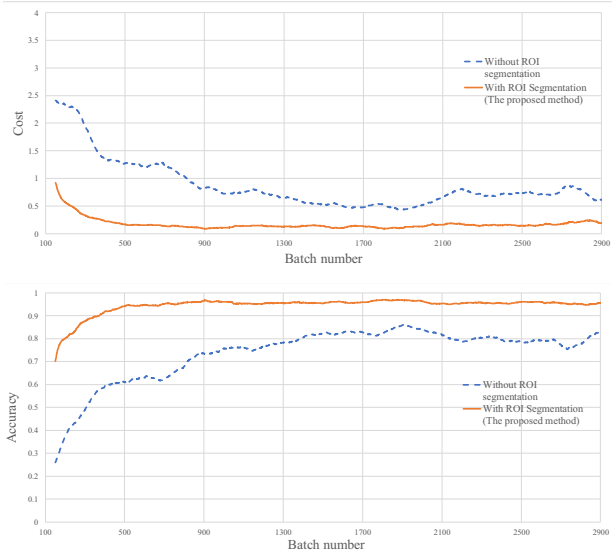

Figure 4. Results of 1.0-m distance test set

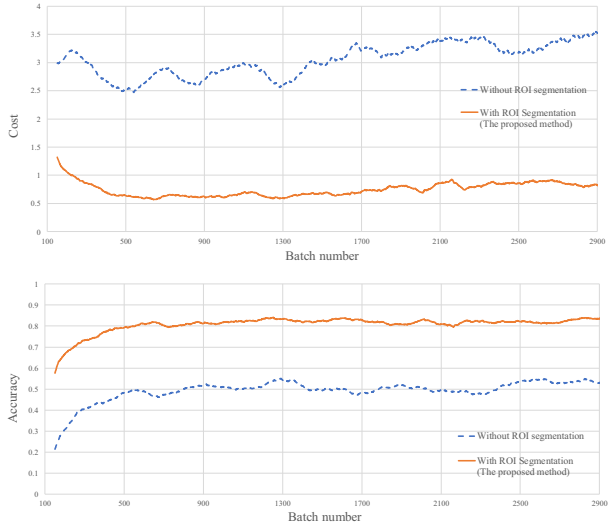

Figure 5. Results of 1.5-m distance test set

With 1.0-m test set, the peak success rates of the methods without and with ROI segmentation are about 84% and 97%, respectively. With 1.5-m test set, they are about 55% and 83%, respectively. Also, the required batch numbers for the cost under 0.5 of the two methods are about 1400 and 200, respectively. This means that the proposed method effectively enhances the success rate and accelerates the training speed since the ROI segmentation has made the problem more clear and simple.

IV. CONCLUSION

The conventional sign language learning system has two distinct drawbacks. The first one is that the area of hand is so

small that the training data should be large. The second one is that it is hard to extend the data set for adding a new gesture to be learned. To solve these problems, we introduced a new sign language learning method that extracts hand area as the ROI before learning, using object detection network. As a result, the success rate as well as the learning speed is distinctly improved. In addition, as it extracts hand area through object detect, it does not need to create lots of data in various situations and it becomes much easier to add a new gesture to be learned.

REFERENCES

[1] Ian Lim, oshua Lu, Claudine Ng, Thomas Ong and Clement Ong, "Sign-language Recognition through Gesture & Movement Analysis (SIGMA)." Computer, Dec. 2015.

[2] Jain, Sanil, and Kadi Vinay Sameer Raja, "Indian Sign Language Gesture Recognition," 2015.

[3] Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, "Sign language recognition and translation with kinect," IEEE Conf. on AFGR, 2013.

[4] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, Benjamin Schrauwen, "Sign language recognition using convolutional neural networks," Workshop at the European Conference on Computer Vision. Springer International Publishing, 2014.

[5] Helen Cooper., Eng-Jon Ong, Nicolas Pugeault, Richard Bowden, "Sign language recognition using sub-units," Journal of Machine Learning Research, Jul. 2012, pp. 2205-2231.

[6] Weinland, Daniel, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," Computer vision and image understanding 104, Feb. 2006, pp. 249-257.

[7] Agarwal, Anant, and Manish K. Thakur, "Sign language recognition using Microsoft Kinect," Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.

[8] Y Ji, S Kim, KB Lee, “Sign Language Learning System with Image Sampling and Convolutional Neural Network.” Robotic Computing (IRC), IEEE 2017

[9] Sahoo, Ashok K. Gouri Sankar Mishra, and Kiran Kumar Ravulakollu, "Sign language recognition: state of the art," ARPN Journal of Engineering and Applied Sciences, vol. 9. Feb. 2014, pp. 116-134.

[10] Jens Forster, Oscar Koller, Christian Oberd¨orfer, Yannick Gweth, Hermann Ney, "Improving continuous sign language recognition: Speech recognition techniques and system design," Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France, Aug. 2013.

[11] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney, "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather." LREC, 2014.

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi,” You only look once: Unified, real-time object detection” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[13] Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert, "Unsupervised Learning using Sequential Verification for Action Recognition," arXiv preprint arXiv:1603-08561, Mar. 2016.

[14] Chéron, Guilhem, Ivan Laptev, and Cordelia Schmid, "P-cnn: Pose-based cnn features for action recognition," Proceedings of the IEEE International Conference on Computer Vision, 2015.

[15] Weinzaepfel, Philippe, Zaid Harchaoui, and Cordelia Schmid, "Learning to track for spatio-temporal action localization." Proceedings of the IEEE International Conference on Computer Vision, 2015.

[16] Srebro, Nathan, and Adi Shraibman, "Rank, trace-norm and max-norm, " International Conference on Computational Learning Theory. Springer Berlin Heidelberg, 2005.

[17] Simard, Patrice Y. David Steinkraus, and John C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," ICDAR. Vol, Mar. 2003.

[18] Jia, Yangqing, et al, "Caffe: Convolutional architecture for fast feature embedding," Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.