Dynamic Hand Gesture Based Sign Word Recognition Using Convolutional Neural Network with Feature Fusion

Md Abdur Rahim[1], Jungpil Shin[2], Md Rashedul Islam[3]

School of Computer Science and Engineering, The University of Aizu
Aizuwakamatsu, Fukushima, Japan
[1]rahim_bds@yahoo.com, [2]jpshin@u-aizu.ac.jp, [3]rashed.cse@gmail.com

Abstract

Gesture-based sign language recognition systems play an important role in human-computer interaction to develop communication between deaf communities and other people. Where the deaf community, hard of hearing, and deaf family members express their feelings and communicate with others. In this case, hand gestures have been a promising subject and applied to the very practical application of sign language recognition (SLR). SLR is highly influenced by the recognition of hand, as the sign word is a form of communicative gesture. However, the diversity and complexities of the gestures of the hand can greatly affect reliability and recognition rates. To solve this problem, this paper introduces an effective sign word recognition system using a deep learning technique, including feature fusion convolutional neural network. In the proposed system, the input image is captured from the live video using a low cost device, such as a webcam and preprocessed hand gesture image. The pre-processing is accomplished with the conversion of YCbCr, binarization, erosion and finally hole fillings. Two channels of CNN are used to extract the features from preprocessed images. The feature fusion is performed at the fully connected layer and this feature is used for gesture classification by the softmax classifier. An experimental setup established in our laboratory environment and the user can recognize the signs of fifteen common words in real-time. The experimental results show high recognition accuracy in gesture-based sign word recognition compared with the state-of-art systems.

Key words: hand gesture, segmentation, sign word, convolutional neural network (CNN)

Introduction

Sign Language Recognition (SLR) is an important medium for communicating with the deaf community and hearing impaired people. Each gesture of the sign language conveys a specific meaning. It has some complexity that the whole language is expressed by the diversity of hand-shaped, body movements and even facial expression. Currently, it is dependent on human-based translation services to increase contact with the deaf community, however, it is problematic and costly as it is related to human skill. Therefore, we focus on hand gesture-based sign word recognition which provides a powerful and reliable interface that allows people to receive immediate feedback from the signer's symptoms as text without the translation services. Hand gestures are a challenging problem that is used to express their thoughts and feelings; this helps to strengthen the information distributed in essential conversation. However, hand tracking and gesture recognition play an important role in human-computer-interaction (HCI) [1-3]. For the advancement of HCI, we can achieve human expression by recognizing the gesture, which can reduce the impassivity among the deaf and common people. In the previous study, many scholars contributed to the development of sign language recognition research. The geometric features such as contour, edges are emphasized to recognize the hand gesture [4-5]. However, image processing is very complicated and takes a lot of time and the recognition accuracy of these approaches is less than our proposed method. Rahim et al. presented an artificial neural network based Bengali sign language recognition [6]. However, the system is only tested from the known dataset and it was not considered light illumination. In [7], the author proposed a skeletal data based hand gesture recognition. It requires huge computation to identify the gesture and palm position. Nowadays, deep learning strategies have achieved the most outstanding results on image recognition and computer vision [8-9]. It has multiple hidden layers which are based on the nonlinear network model that can achieve sophisticated accuracy. Xiao Yan Wu [10] introduced a novel approach to recognize the hand gesture using the two input channel of CNN. A review of sophisticated technology for the recognition of hand gesture and sign languages described in [11]. The authors described different steps such as pre-processing, segmentation, feature extraction and classification are implemented to detect hand gestures and sign language recognition. To improve communication with the deaf community and the common people, we propose a hand gesture based human-computer interface in this paper. A deep learning technique such as CNN is used to analyze the images and extracts different features in the images. The feature fusions are completed at the fully connected layer and the softmax classifier is used to classify the recognition results.

We organized this paper as follows. Section 2, we explain the workflow and proposed a model. The experimental results and analysis of this system are described in Section 3. Section 4 declares research outcome and future plans.

Proposed Methodology

Fig. 1 represents the proposed methodology of hand gesture based isolated sign word recognition system. The input image is captured from the region of interest (ROI) of live video frames using a webcam. The overall procedure of hand gesture recognition is described below.

A. Hand Gesture Segmentation

In order to segment hand gestures from input images, a data

ISBN: 978-1-7281-0110-1

pre-processing techniques are employed. We convert the grayscale image to an input image by transforming the YCbCr from RGB color space. The YCbCr has the luminance (Y) and chrominance (Cb and Cr) color values. The pixel values of the grayscale image are between 0 and 255, 0 are generally black and 255 is white. As 128 sets a threshold value and the pixel values are redefined as 0-127 to 0 and 128-255 to 255, we can process the image in grayscale with binary images. Then we apply erosion to the binary image which removes the regions of boundaries of foreground pixels. Thus the size of the foreground pixels is shrunk in size, and the holes in the area became larger. Finally, we fill those holes and accept the image of the hole that is used to extract the feature. Fig. 2 shows the preprocessing steps of an input image.

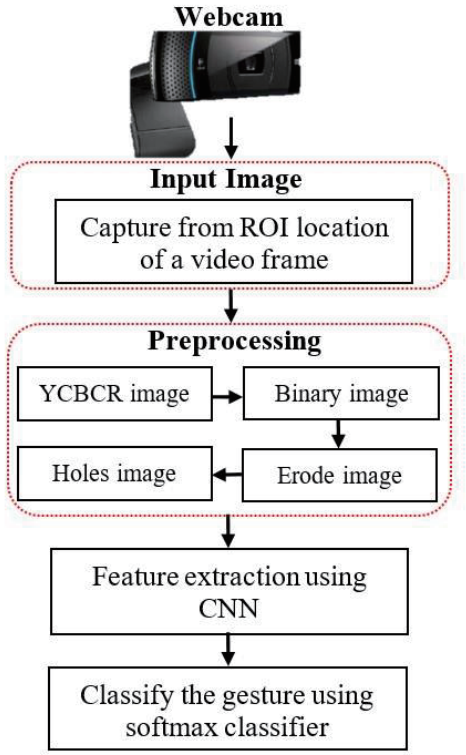

Fig. 1 Overall architecture of sign word recognition system.

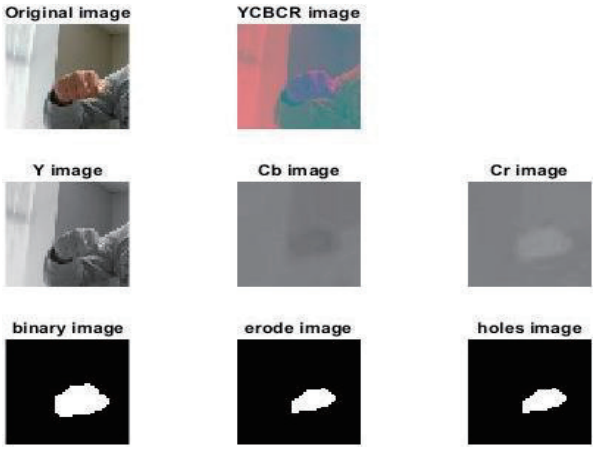

Fig. 2 Preprocessing steps of an input image.

C. Feature Extraction and Classification

As a popular class of machine learning techniques, the Convolutional Neural Network (CNN) has expanded significantly in the technological advances of human-computer interaction [12]. The CNN described in Fig. 3 includes convolutional, pooling and fully connection layer, activation function, and the classifier. Gesture images and segmented images are the input of two channels. The convolutional level is to detect the input local features and move the convolution through the specific length of kernel steps. Since there is no difference in the convolution kernel, but the weight parameters are obtained through the running level by level. In pooling layer, the transformation of the function takes place in the overlapping field of the output level of convolutional. Therefore, a higher level of invariant features can be achieved. However, the pooling layer can reduce the data layer while saving feature information. Generally, the fully connected layer connects to the latest level of pooling and classifier, which is used to categorize various features expressed by multiple features. The proposed method is used by the feature descriptor to provide complete information on the last fully connected layer. The fully connected layer can only accept one-dimensional data, where we used Flatten function. Finally, the fully connected layer integrates all the features and provides to the softmax classifier. The Relu function is used as an activation function.

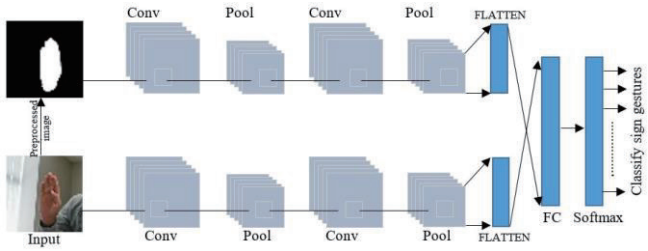

Fig. 3 Proposed Model.

Experimental Results

A comprehensive test is managed to recognize different sign gestures with the proposed architecture. This section describes datasets and system setups as well as the proposed architectures that have been implemented.

A. Experimental Dataset

In the examination, hand gesture-based sign word images are used to evaluate the effectiveness of the proposed method. The dataset considers fifteen isolated gestures to collect images using a webcam. For each gesture, 900 images are captured. Therefore, a total of 13,500 images are used to sign word recognition. For creating a gesture database, three volunteers are performed. We collected 300 images for each gesture from each individual. The gestures images are 200x200 in size. The input images are preprocessed using our proposed methods and provided for feature extraction. Fig. 4 shows the symbol of gesture images of sign word and Fig. 5 depicts the example of segmented gesture images. The experiment was conducted on a computer (Intel Core i5-2400, 3.10 GHz) and GPU GTX 1080 Ti, and the webcam mounted at an appropriate place.

B. Experimental Evaluation

We evaluate gesture-based different sign word recognition in this section. The proposed CNN architecture trained the entire dataset. We used 80% datasets for training and the remaining 20% for testing. After training, the extracted features

ISBN: 978-1-7281-0110-1

are transferred to the classification process. However, the average accuracy of sign word recognition is 96.96%. The system achieves 100% acceptance of "Fine", "Call" and "Ok" gestures. Fig. 6 shows the confusion matrix of recognition accuracy. In comparison, we applied the method of reducing the feature dimension [13] and calibrating skin models [14] for training and testing purposes, respectively. The comparison of recognition accuracy is shown in Fig. 7. Table I shows the comparison of gesture recognition accuracy with state-of-art methods. According to the experimental results, our proposed methods achieve better results than the state-of-art methods because of using efficient segmentation technique that helps identify the area of the hand properly. Moreover, feature fusion of CNN extracts efficient features which improves the classification accuracy.

Fig. 4 Samples of sign word dataset images.

Fig. 5 Example of the preprocessed images of different sign gestures.

For real-time performance, the preprocessed images and auxiliary features are presented by vectors in pre-trained networks, and the gesture is classified according to the most active output results. Thus the input image can also be processed with the proposed architecture and the total time required for the whole process is 23.03 ms, which is enough to recognize in real-time. In fact, the image is available at 33.33 ms per frame in webcam. Therefore, we evaluated the recognition accuracy of different sign gestures among different users. Fifteen respondents were asked to perform the gesture of sign word, requesting users to execute ten times of each gesture. Fig. 8 presents the simulation performance of one user. Fig. 9 show the average recognition accuracy of all users of fifteen sign gestures.

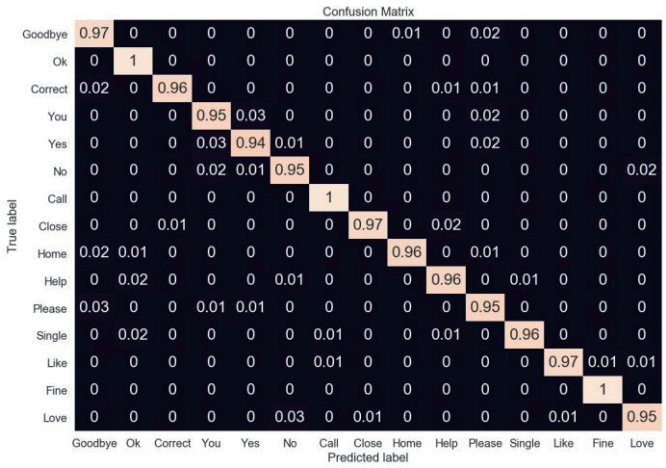

Fig. 6 Average recognition accuracy of different sign word of all users.

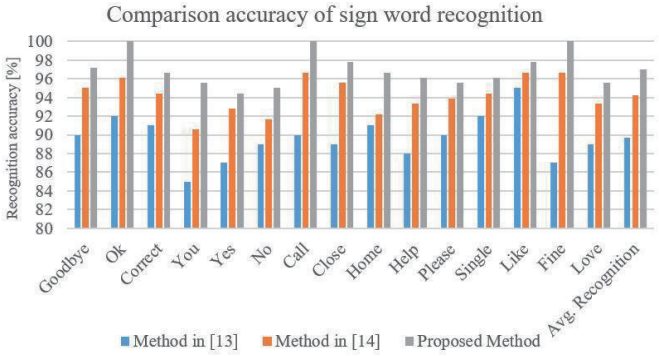

Fig. 7 Comparison of recognition accuracy of isolated sign word.

TABLE I
COMPARISON OF GESTURE RECOGNITION ACCURACY

Model	Method	Function	Gestures	Accuracy
Ref. [13]	DWT & SVM	Hand Gestures	15 Gestures	89.67% (Evaluated using our dataset)
Ref. [14]	CNN	Hand Gestures	7 Gestures	95.96% (From [14])
Ref. [14]	CNN	Hand Gestures	15 Gestures	94.22% (Evaluated using our dataset)
Proposed	Segmentation & CNN	Sign Word	15 Gestures	96.96%

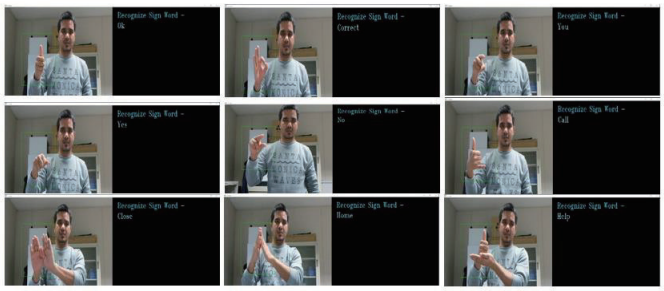

Fig. 8 Simulation of sign word recognition of a user.

ISBN: 978-1-7281-0110-1

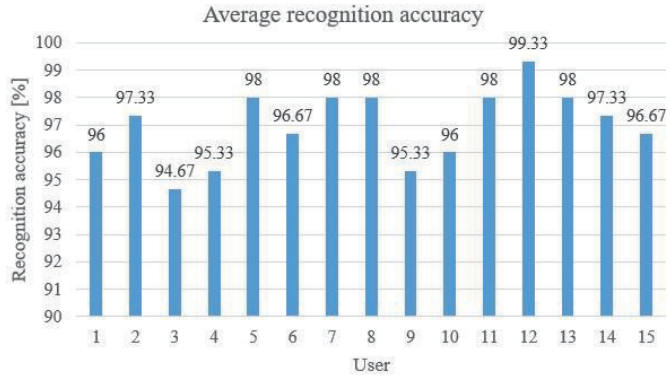

Fig. 9 Average recognition accuracy of all users.

Conclusion

In this paper, we developed a dynamic hand gesture based sign word recognition system using a convolutional neural network that translates sign gestures into text. The different steps such as preprocessing, feature extraction and classification are implemented to detect hand gestures and sign word recognition. The hand gesture images are preprocessed according to the YCbCr conversion, grayscale image selection, binarization, erosion and fills the gaps. A deep learning technique i.e., CNN with feature fusion are used to extract the features and provide to the classifier for classification. A softmax classifier is used to classify these gestures. Also, the gesture-based sign-word recognition system is implemented by a webcam in real-time. As a result, the average acceptance of gesture-based sign word recognition is 96.96% which leads to better results than state-of-art systems.

The future work will be aimed at developing the hand gesture based sign words and novel sentences interpretation algorithm for sign language recognition system.

References

[1] F.S. Chen, C.M. Fu and C.L. Huang. "Hand gesture recognition using a real-time tracking method and hidden Markov models." Image and vision computing, vol. 21, no. 8, pp. 745-758, Aug. 2003.

[2] G. Marin, F. Dominio, and P. Zanuttigh. "Hand gesture recognition with jointly calibrated leap motion and depth sensor." Multimedia Tools and Applications, vol. 75, no. 22, pp. 14991-15015, Nov. 2016.

[3] P. Kumar, H. Gauba, P.P. Roy, and D.P. Dogra. "Coupled HMM-based multi-sensor data fusion for sign language recognition." Pattern Recognition Letters, vol. 86, pp. 1-8, Jan. 2017.

[4] D. Lifeng, R. Jun, M. Qiushi, W. Lei, "The gesture identification based on invariant moments and SVM[J]." Microcomputer and Its Applications, vol. 31, no. 6, pp. 32-35, 2012.

[5] Y.H. Sui, Y.S. Guo, "Hand gesture recognition based on combing Hu moments and BoF-SURF support vector machine." Application Research of Computers, vol. 31, no. 3, pp. 953–956, 2014.

[6] M.A. Rahim, T. Wahid, M.K. Islam, "Visual recognition of Bengali sign language using artificial neural network." International Journal of Computer Applications, vol. 94, no. 17, Jan. 2014.

[7] M.A. Rahim, J. Shin, and M.R. Islam, "Human-Machine Interaction based on Hand Gesture Recognition using Skeleton Information of Kinect Sensor." In Proceedings of the 3rd International Conference on Applications in Information Technology, ACM, pp. 75-79, Nov. 2018.

[8] T. Yamashita, T. Watasue, "Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning." IEEE international conference on image processing (ICIP), pp 853–857, Oct. 2014.

[9] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks." IEEE Access, 2019.

[10] X.Y. Wu. "A hand gesture recognition algorithm based on DC-CNN." Multimedia Tools and Applications, pp. 1-13, 2019.

[11] M.J. Cheok, Z. Omar, and M.H. Jaward. "A review of hand gesture and sign language recognition techniques." International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, pp. 131-153, Jan. 2019.

[12] S.F. Chevtchenko, R.F. Vale, V. Macario, and F.R. Cordeiro. "A convolutional neural network with feature fusion for real-time hand posture recognition." Applied Soft Computing, vol. 73 pp. 748-766, Dec. 2018.

[13] R.A. Bhuiyan, A.K. Tushar, A. Ashiquzzaman, J. Shin, and M.R. Islam, "December. Reduction of gesture feature dimension for improving the hand gesture recognition performance of numerical sign language." IEEE 20th International Conference of Computer and Information Technology (ICCIT), pp. 1-6, Dec. 2017.

[14] H.I. Lin, M.H. Hsu, W.K. Chen, "Human hand gesture recognition using a convolution neural network." In IEEE International Conference on Automation Science and Engineering (CASE), pp. 1038-1043, Aug. 2014.

ISBN: 978-1-7281-0110-1