2018 4th International Conference on Frontiers of Signal Processing

Hand Gesture Feature Extraction Using Deep Convolutional Neural Network for Recognizing American Sign Language

Md Rashedul Islam
School of Computer Science and Engineering
University of Aizu
Fukushima, Japan
e-mail: rashed.cse@gmail.com

Ummey Kulsum Mitu
Department of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
e-mail: ummey.kulsum@gmail.com

Rasel Ahmed Bhuiyan
Department of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
e-mail: raselcse34@gmail.com

Jungpil Shin
School of Computer Science and Engineering
University of Aizu
Fukushima, Japan
e-mail: jpshin@u-aizu.ac.jp

***Abstract*—In this era, Human-Computer Interaction (HCI) is a fascinating field about the interaction between humans and computers. Interacting with computers, human Hand Gesture Recognition (HGR) is the most significant way and the major part of HCI. Extracting features and detecting hand gesture from inputted color videos is more challenging because of the huge variation in the hands. For resolving this issue, this paper introduces an effective HGR system for low-cost color video using webcam. In this proposed model, Deep Convolutional Neural Network (DCNN) is used for extracting efficient hand features to recognize the American Sign Language (ASL) using hand gestures. Finally, the Multi-class Support Vector Machine (MCSVM) is used for identifying the hand sign, where CNN extracted features are used to train up the machine. Distinct person hand gesture is used for validation in this paper. The proposed model shows satisfactory performance in terms of classification accuracy, i.e., 94.57%**

Keywords-human-computer-interaction (HCI); convolutional neural network (CNN); Hand Gesture Recognition; sign language; multi-class support vector machine (MCSVM)

I. INTRODUCTION

Human-computer-interaction referred to as HCI is an interacting interface between humans (users) and machines (computers). Through HCI, humans and computers interact with each other in a novel way. Nowadays, it's a fascinating research field, which is focused on the designs and uses of computer technology and most particularly, the interacting interfaces between humans and machines. The HCI technology has been remarkably expanded and raised up with the changes in technology [1].

Conventionally in HCI, the command line interface (CLI) uses the keyboard and the graphical user interface (GUI) uses a keyboard and a mouse along with graphics to provide an interface for humans to interact with computers. On the basis of effective usability, new technologies introduce new user interfaces like Direct Neural Interfaces (DNI) in HCI [2]. Non-touch, gesture, and voice interface are becoming popular no a day. Hence, DNI is a new technology of HCI to communicate with a machine by recognizing the brain signal without any physical participation.

DNIS is often directed at assisting, augmenting, or repairing human cognitive functions. But, in every kind of applications, this technology is difficult and expensive to embed. So, those newly added technologies are introduced as adaptive to the real applications based on the requirements and cost-effectiveness. Whatever, those newly introduced technologies can't reach the satisfaction level of users significantly yet. To overcome the challenges, many researchers working on improving those interfaces at the level of effectiveness, usability, and robustness [3].

An ideal interface should have some common features criteria like usability, accuracy, affordability, and scalability. Nowadays, the Human gesture has become a popular HCI interface, and the usage of human gesture is increasing rapidly, which meets all these criteria.

HGR has lots of applications in different fields such as computer game, virtual reality and sign language recognition (SLR). Among them, SLR is the most used technique where vocal transmission is impossible. Disable people should have the capability to recognize sign generated by others. Therefore, many researchers have taken a challenge to present an assembler prototype for the American Sign Language (ASL).

Several types of research have been done on human sign recognition with a few numbers of symbols. However, sign recognition for alphabet is more challenging. Many researchers invented approaches related to human body and hand gesture to enhance the usage of technology. Kilioz et al. introduced an effective approach for recognizing dynamic hang gesture on the basis of real-time HCI [4]. Modanwal et al. solved the gap between machine and blind people by introducing gesture recognition [5].

978-1-5386-7853-4/18/$31.00 ©2018 IEEE

Rempel et al. worked for understanding sign language using a human hand gesture [6]. Denkowski et al. proposed a model to control residential and commercial building components using human gesture in a natural way [7]. Liang et al. used a hidden Markov Model (HMM) for recognizing the sign language [8]. Because of a large number of gesture of alphabets, those models show sub-optimal results for alphabet sign recognition.

From this point of view, this paper proposes an efficient feature extraction process using Convolutional Neural Network (CNN). The CNN consists of one or more fully connected convolutional layers as standard multilayer neural network [9]. CNN architecture is designed for handling 2D images efficiently [10]. Also, CNN has several dynamic parameters to train up the machine easily [11]. Finally, the Multiclass Support Vector Machine (SVM) is used for recognizing gesture for sign language.

The rest of the paper organized is as follows. Section II describes the different parts of the proposed model. The experimental result is discussed in section III. Finally, section IV concludes this paper.

II. PROPOSED MODEL

Identify the ASL alphabet depending on human hand gesture is the basic idea of our proposed model. The working procedure of the proposed model is shown in Figure 1.

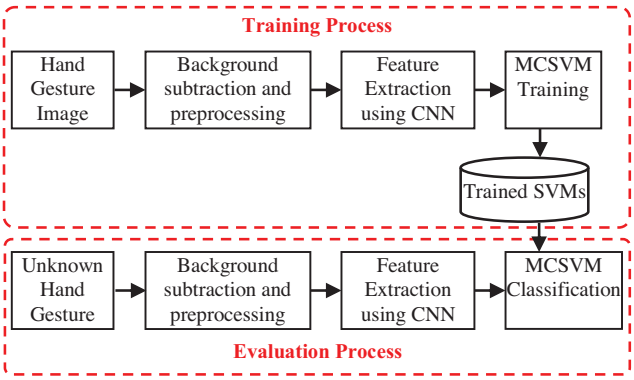

Figure 1. Working procedure of the proposed model.

A. Experimental Setup and Preprocessing Gesture Image

In order to capture a video frame from the webcam, the experimental setup has been established at the very beginning. To discard the unwanted area of the video frame, a particular area is fixed as a Region of Interest (ROI). Figure 2 shows the Region of Interest.

In the background subtraction process, firstly, a frame of background picture is taken without human hand gesture. The captured frame is subtracted for the video frame of hand gesture for getting the hand sign image. Because of the background process and light effect, there are some noises in gesture images. To reduce the noise from the image, a median filter is used. Then the image is converted into a grayscale image. Finally, the human hand gesture images are taken for 26 alphabet sign of ASL for three different persons. For each sign, there are 120 images and for each person, there are 3,120 (26x120) images. Figure 3 shows the preprocessing of gesture images.

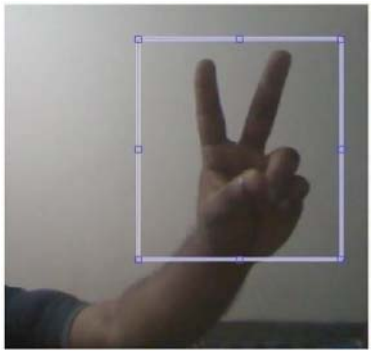

Figure 2. The Region of Interest (ROI).

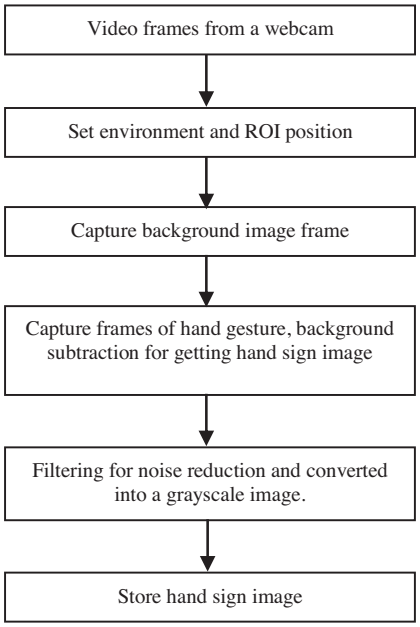

Figure 3. Preprocessing of gesture images.

B. Gesture Feature Extraction Using CNN

In this proposed feature extraction model, the feature vector is extracted from a video frame using Deep Convolutional Neural Network (DCNN). All the extracted feature values of the images are stored in a file after extraction.

There are many effective machine learning algorithms for feature extraction. Convolutional Neural Network is one of the best techniques in the field of deep learning. CNN can be used on a large scale of diverse images. For a wide range of images, CNN can extract potential features for the classification model.

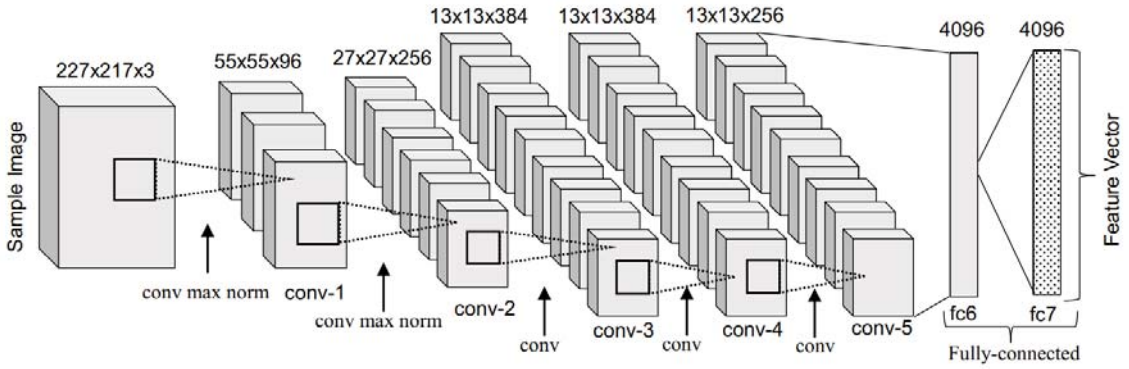

Figure 4. The architecture of proposed feature extraction using CNN.

There are several networks of CNN. The "AlexNet" is one of the most popular network for its effectiveness. In AlexNet network, there are five convolution layers and three fully-connected layers in the network [9]. Input dimensions are defined in the first layer. In this paper, we used the input image size is 227-by-217-by-3. Make up the bulk of CNN by intermediate layers. Figure 4 shows the architecture of Convolutional Neural Network used in this proposed model.

CNN produces an activations method, is used as an input image for each layer. The convolution process is running layer by layer. However, for image feature extraction, there are only a few layers that are suitable within a CNN. In this proposed model, 'fc7' layer is considered for feature extraction. Using this layer, basic image features are captured by the layers at the beginning of the network. Deeper network layers are processed these primitive features and combine the early features to form higher level image features. All of these higher-level features are well suited for the tasks of classification. Because deeper network layers are combined all of these primitive features into a richer image representation [10].

C. Gesture Classification Using SVM

Lastly, we used non-linear MCSVM for the classification of each and every alphabetic sign in the last section of this proposed model. SVM is a mostly used learning method for the purposes of classification of extracted features and regression. As SVM is a binary classifier basis on the supervised learning approach for classifying data into two classes by drawing a hyperplane [12].

The core working procedure of SVM is about classifying the inputted sample data set into two distinct classes using a hyperplane. Many datasets are not linear in that case; hyperplane is unable to classify those datasets into two classes. Kernel function successfully concludes this problem of classifying non-linear datasets [13]. Gaussian radial basis function, Polynomial function, and hyperbolic tangent function are some common kinds of non-linear kernel functions. Among them, Gaussian radial basis kernel function is most used non-linear kernel function. In this paper, the Gaussian radial basis kernel function used, which can be expressed by equation (1).

$$k(sv_i, sv_j) = \exp\left(\frac{\left\|sv_i, sv_j\right\|^2}{2\delta^2}\right) \quad (1)$$

where k is the processing function of two separate input parameter sv_i and sv_j. Another independent variable is needed to process inputted parameters or feature vector for finding the width of effective basis kernel function which can be denoted as δ.

Generally, the SVM is a binary classifier. However, there is some basic form of SVM like - one-against-one (OAO), one-against-all (OAA), one-acyclic-graph (OAG) etc. [14]. From those methods, OAA is used in this proposed model due to the least complications of this multi-class non-linear classifying method. The OAA-MCSVM contains total twenty-six SVMs working as a parallel way as follows in Figure 5. In each SVM, one class is differentiated from others and the concluding decision is taken from this process by selecting the largest outputted value's SVM.

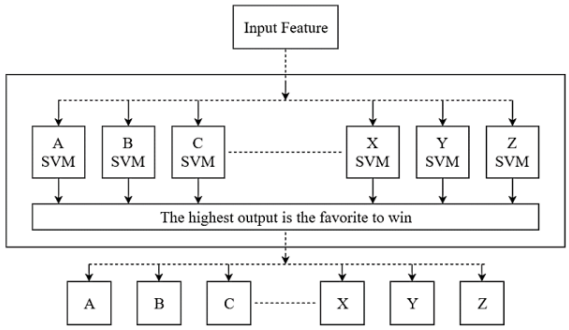

Figure 5. An OAA-MCSVM structure for identifying alphabets.

III. EXPERIMENTAL RESULT AND EVALUATION

The proposed model is evaluated by a constructed dataset which contains 26 signs of three different persons. Every sign contains 120 images for each person. Therefore, there are 9,360 (3x26x120) images in total. The whole dataset is branched into two sets. First branched set contains 30% images for training and another one contain rest of the 70% images for testing. Figure 6 shows the alphabet sign of ASL.

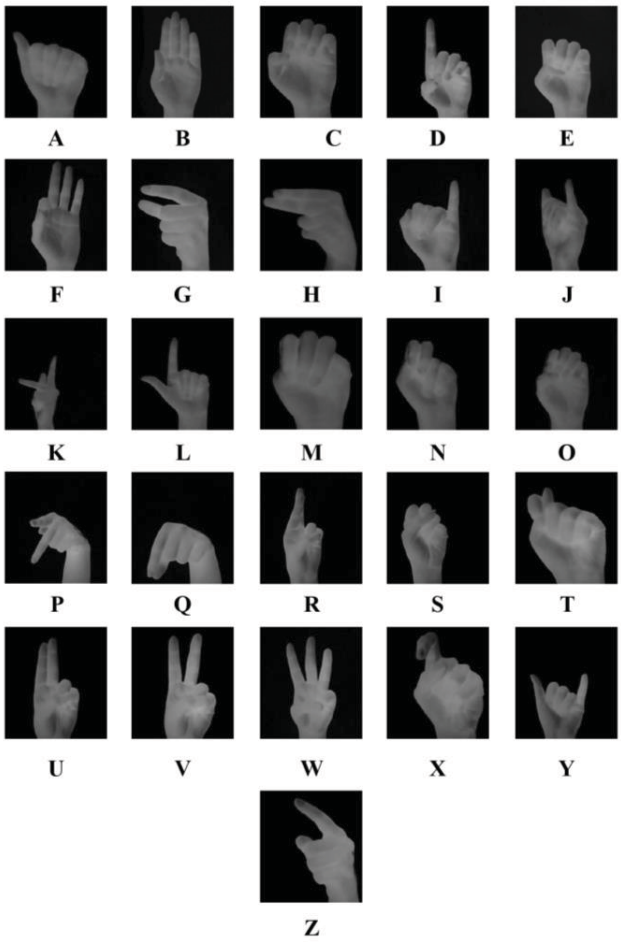

Figure 6. ASL representation of alphabet.

The convolutional neural network is used for feature extraction. After feature extraction of the images using CNN, therefore a 4096x2808 number of features for training and 6552x4096 number of features for testing we find. All of these features are informative, which helps to classify the categories of each person sign.

Using those more informative training and testing features, classified each sign of each person performed by the MCSVM. The classification accuracy obtained satisfactorily which is 94.57%. The accuracies of individually sign are shown in Table I. The person wise results of our proposed model is shown in Table II.

IV. CONCLUSION

A significant challenge in real-life applications is hand gesture recognition in terms of the accuracy and robustness associated with it. Non-touch hand gesture recognition of ASL is presented in this paper, the input gestures are collected using the webcam. At the very beginning, the still hand image frame captured from a running video frame and performed DCNN in order to find more informative features. Finally, identify the alphabet sign using MCSVM. For validation of this proposed model, our constructed dataset according to ASL conventions is used. The classification accuracy obtained 94.57% which is significant for introducing the SLR of ASL for disable people as an output of HCI.

TABLE I. SIGN WISE CLASSIFICATION ACCURACY

Sign	Recognition Accuracy	Average Accuracy
A	88.49%	94.57%
B	100%	
C	91.67%	
D	99.60%	
E	81.35%	
F	98.02%	
G	99.21%	
H	98.81%	
I	99.60%	
J	95.63%	
K	92.46%	
L	100%	
M	87.70%	
N	86.90%	
O	82.94%	
P	92.46%	
Q	99.60%	
R	96.03%	
S	98.81%	
T	94.44%	
U	94.05%	
V	96.83%	
W	97.62%	
X	96.43%	
Y	98.02%	
Z	92.06%	

TABLE II. PERSON WISE CLASSIFICATION ACCURACY

Person	**Recognition Accuracy**	**Average Accuracy**
P1	93.19%	94.57%
P2	95.29%	
P3	95.24%	

REFERENCES

[1] A. Dix, "Human-computer interaction," in Encyclopedia of Database Systems, Springer US, pp. 1327–1331, 2016.

[2] K. Nandakumar and J. L. Funk, "Understanding the timing of economic feasibility: The case of input interfaces for human-computer interaction," Technology in Society, vol. 43, pp. 33 – 49, Nov. 2015.

[3] B. Laurel and S. J. Mountford, "The Art of Human-Computer Interface Design," Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1990.

[4] N. C. Kiliboz and U. Gudukbay, "A hand gesture recognition technique for human computer interaction," Journal of Visual Communication and Image Representation, vol. 28, pp. 97 – 104, Apr. 2015.

[5] G. Modanwal and K. Sarawadekar, "Towards hand gesture based writing support system for blinds," Pattern Recognition, vol. 57, pp. 50 – 60, Sep. 2016.

[6] D. Rempel, M. J. Camilleri, and D. L. Lee, "The design of hand gestures for human-computer interaction: Lessons from sign language interpreters," International Journal of Human-Computer Studies, vol. 72, no. 10-11, pp. 728 – 735, Oct.-Nov. 2014.

[7] M. Denkowski, K. Dmitruk, and L. Sadkowski, "Building automation control system driven by gestures," in Proceedings of the 13th IFAC

and IEEE Conference on Programmable Devices and Embedded Systems, vol. 48, no. 4, pp. 246 – 251, 2015.

[8] R.-h. Liang and M. Ouhyoung, "A sign language recognition system using hidden markov model and context sensitive search," in Proceedings of the ACM Symposium on Virtual Reality Software and Technology. Hongkong, pp. 59–66, 1996.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 1097–1105, Dec. 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computer Vision and Pattern Recognition, pp. 1409–1556, 2014.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in Proceedings of the 31st International Conference on Machine Learning, vol. 32, no. 1, pp. 647–655, Jun 2014.

[12] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004, vol. 3, pp. 32–36, 2004.

[13] M. R. Islam, J. Uddin, and J.-M. Kim, "Acoustic Emission Sensor Network Based Fault Diagnosis of Induction Motors Using a Gabor Filter and Multiclass Support Vector Machines." Adhoc & Sensor Wireless Networks, vol. 34, pp. 273-287, Dec. 2016.

[14] J. Manikandan and B. Venkataramani, "Evaluation of multiclass support vector machine classifiers using optimum threshold-based pruning technique," IET signal processing, vol. 5, no. 5, pp. 506–513, 2011.