Hand Gesture Recognition using PCA based Deep CNN Reduced Features and SVM classifier

Jaya Prakash Sahoo, Samit Ari
Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela
Odisha, 769008, India
Email: sahoo.jprakash@gmail.com, samit@nitrkl.ac.in

Sarat Kumar Patra
Indian Institute of Information Technology, Vadodara
Gandhinagar, Gujarat, 382028, India
Email: skpatra@iiitvadodara.ac.in

***Abstract*—Automatic recognition of vision based static hand gesture images is a challenging task due to illumination changes, diversity in user hand shape and high inter class similarities. This paper proposes novel techniques to develop a user independent hand gesture recognition system, considering the above challenges. First, the gesture recognition performance is analyzed using proposed pre-trained AlexNet features. In this proposition the deep features are extracted from fully connected (FC) layers such as 'FC6 and 'FC7' of pre-trained AlexNet. A support vector machine (SVM) based classifier with linear kernel is used to classify gesture poses. The highest recognition accuracy is evaluated using the deep feature extracted from 'FC6 and 'FC7' independently and combination of both the feature vector with SVM classifier. Second, feature dimension of deep features are reduced using principal component analysis (PCA) based dimension reduction technique for further improvement in gesture recognition accuracy. The performance of the proposed technique is evaluated using leave-one-subject-out cross validation (LOO CV) and holdout CV test. The extensive analysis is performed on 36 American Sign Language (ASL) benchmark static hand gesture dataset using both the CV test. The experimental result shows that, the proposed technique is superior as compared to state-of-the-art techniques.**

***Index Terms*—Hand gesture recognition, pre-trained CNN, deep CNN feature extraction, PCA, support vector machine.**

I. Introduction

Hand gesture recognition has been one of the active research area in the field of human-computer interface (HCI) due to its flexibility and user friendly [1]. Several applications on hand gesture recognition reported in the literature are, sign language recognition [2], writing support system for blind people [3], control of a software interface [4] ans so on. The use of hand gesture based system in public places such as ATM, railway station for booking of platform ticket etc. where, the gesture recognition is greatly affected by the variation in light intensity and also due to daylight. In such case the segmentation of hand region is largely affected by intensity of light. Therefore, a robust and user independent hand gesture recognition system is required which can work in such variation of light intensity. Although gesture recognition system is possible by depth camera but our objective is to develop a vision hand gesture recognition system with a low cost camera like webcam. Also in depth based hand gesture recognition system, lots of debris are appeared as noise in the depth image due to very low illumination [5].

Hand gesture recognition system is developed using the following major steps such as hand segmentation, feature extraction and recognition of gesture pose using the extracted features [1]. A novel wristband-based contour features (WBCFs) was proposed by Lee *et al.* [6], for recognition of 29 Turkish finger spelling sign hand gestures. The segmentation of hand is obtained by detecting black wristbands on both the hands and using skin color based filtering technique. Then the gestures are recognized using the WBCFs feature and a matching algorithm. A light invariant hand gesture recognition technique was developed by Chaudhary *et al.* [7] using orientation histogram feature and artificial neural network (ANN) classifier. Six different hand gestures with variation in skin color and light intensity were collected from online search and manually to develop the system and a overall accuracy of 92.86% was achieved using the proposed technique. Recognition of 36 American Sign Languages (ASL) of Massey University (MU) dataset using fusion of features such as Fourier descriptor (FD), Zernike moments (ZM), Hu moments, complex moments and Gabor features (GB) are analyzed by Chevtchenko *et al.* [2] using artificial neural network classifier. Then a multi-objective genetic algorithm was used to optimized the feature vector for accurate recognition of hand postures. Recognition accuracy up to 97.63% was achieved on combination of GB-ZM features with holdout cross validation (CV) test. Fusion of four hand crafted features with the convolutional neural network (CNN) at last fully-connected layers was proposed by Chevtchenko *et al.* [8] for the recognition of ASL sign language. Recognition accuracy of 84.02% and 98.05% was achieved for leave-one-subject-out cross validation (LOO CV) and holdout CV respectively, on MU dataset. From the above literature survey, it is found that the gesture recognition performance is affected by the following problems such as hand segmentation, illumination variation, similarities between the gesture poses and testing the model with different users.

This paper demonstrates the gesture recognition technique using novel deep CNN feature and optimization of feature dimension for fast and accurate recognition of hand gesture. The contribution of our proposed work are as follows:1) Analysis of gesture recognition performance using deep features extracted from fully connected layers such as 'FC6' and 'FC7'

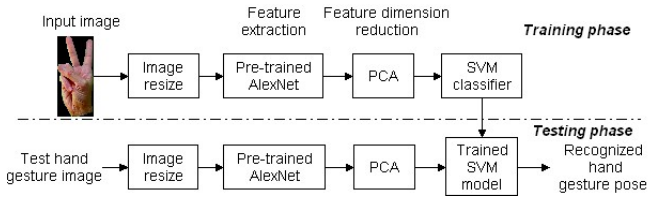


Fig. 1. Block diagram representation of the proposed hand gesture recognition system.

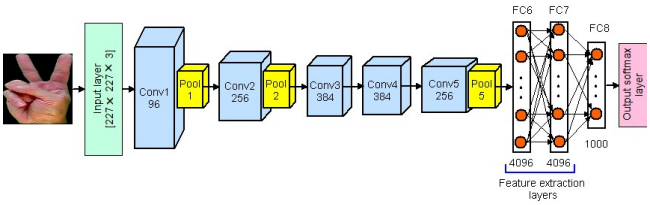


Fig. 2. Architecture of pre-trained AlexNet. Here, Conv: Convolution layer, Pool: Max pooling layer, and FC: Fully connected layer.

of pre-trained AlexNet and combination of both features, 2) Principal component analysis (PCA) based feature dimension reduction of deap features to enhance the performance of gesture recognition accuracy suing SVM classifier, 3) The performance of recognition accuracy is evaluated on 36 gesture pose Msssey University dataset.

The rest of the paper is organized as follows. Methodology to recognized the gesture poses using proposed technique is presented in section II. Challenges in the benchmark dataset and data analysis using validation techniques are described in the section III. The performance of the proposed technique are shown in section IV. Finally, the section V concludes the proposed work.

II. Methodology

The development of hand gesture recognition system using proposed technique is shown in Fig. 1. The gesture recognition system is developed using two phases such as training and testing phase. A training model is developed from input images using the steps as follows: image resize, feature extraction from pre-trained AlexNet model, feature dimension reduction using PCA and development of trained model using SVM classifier. In the test phase, a test gesture sample is applied to recognize the ASL gesture pose using the trained SVM model.

A. Image resize

The input training images are resized according to input image layer size of pre-trained AlexNet model. As the input image layer sizes of pre-trained AlexNet is of resolution [227$\times$ 227 $\times$ 3] so, all the training and testing images are resized according to the above image resolution.

B. Feature extraction from pre-trained CNN model

Pre-trained CNN model is the model trained on large label ImageNet dataset to solve the image category classification problem. These models are generally used to solve the same image classification problem but using different image dataset. In this work, pre-trained AlexNet [9] is used for classification of ASL gesture pose in the hand gesture images. This deep CNN was developed using five convolution layers, three max pooling layers, and three fully-connected layers. The layer architecture of this deep CNN is shown in the Fig. 2.

Convolution layers: In this layer, the features of input image are extracted using convolution operation between a filter or kernel, with the input gesture image. The filter moves on the whole input image to perform the convolution operation with each pixel of input image and generate a feature map. The number of feature maps generated is equal to number of filters used in a convolution layer. As shown in Fig. 3b, out of 96 feature map in first convolution layer (Conv 1) of pre-trained AlexNet, 16 feature maps are presented here, for better visualization. The feature of input gesture image after passing through each feature map are shown in Fig. 3c.

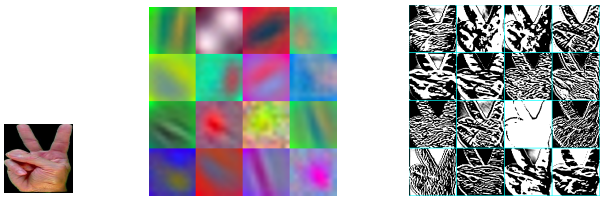
Max-pooling layers: In this layer, the output image of convolution layer is divided in to 2×2 or 3×3 block regions and the maximum value of the region is selected. The advantage of pooling is to down sample the spatial size of each feature maps and also reduces the number of computation in the network.

Fully-connected layers: In this layer, the information extracted by each filters in the last pooling layer are converted into 1D feature vector. Then 1D features are represented as each neuron in the fully connected layer and multiplied with the weight of the neuron to produce the output.

Output softmax layer: This function takes the output of last fully connected layer and transform the real value into a probability distribution between (0,1).

C. Dimension reduction of deep features using PCA

In the feature vector of higher dimensional data the redundant features are having low variance and undesired [10]. Therefore, in dimension reduction technique the objective is to discard those redundant features from the higher dimensional data while keeping the desired information in the reduced feature vector. The advantages of dimension reduction technique are i) it reduces the storage space required at the time of training, ii) the training algorithm became faster, iii) removes redundant features and noises in the data. Here, principal



(a) (b) (c)

Fig. 3. Visualization of learned convolutional feature from first covolutional layer of pre-trained AlexNet. (a) input gesture pose '2', (b) 16 feature map of Conv 1, (c) visualization of features of input image on 16 feature maps

component analysis (PCA) is used to reduced the dimension of deep CNN feature as shown in Fig. 1. The final features are normalized according to zero mean and unit variance before applying to SVM classifier.

D. Training using SVM

The support vector machine (SVM) [11] based classifier is used here, to recognized the gesture pose in the hand gesture image. For a given labeled training images, SVM finds optimal hyperplane between two different classes so that the test gesture images can be categories properly. In this work, one-against-all (OAA) multi-class SVM [11] technique is used to classify the gesture poses. In OAA SVM technique if the dataset is having N output classes then N models are developed using the training data.

III. Experimental Evaluation

A. Benchmark Dataset

MU dataset [12] was developed from 5 subjects with 36 gesture poses (10 ASL digit and 26 ASL alphabets). The challenges in this dataset are variation of illumination in five different directions such as left, right, top, bottom and diffuse, the gesture poses are performed with variation in hand rotation, scale and articulation etc as shown in Fig. 4.

B. Data analysis using validation technique

Two cross validation (CV) methods such as Leave-one-subject-out (LOO) CV [8] and holdout CV [2] tests are used to evaluate the performance of proposed hand gesture recognition technique.

a) LOO CV test: If a dataset is having M subjects then the classifier is trained on the gesture samples of $M-1$ subject and the recognition accuracy is evaluated on the gesture samples of remaining one subject. This CV test is repeated for M times and the mean accuracy is evaluated for the dataset. This is an unbiased CV test as the trained classifier is tested with the gesture samples from a new subject.

b) Holdout CV test: In this CV test, out of the total gesture samples in the dataset, 80% samples are randomly selected to trained the model and the performance of the training model is tested on 20% of remaining gesture samples. Ten observations are repeated for this CV test to measured the performance in terms of mean accuracy.

(a) (b)

Fig. 4. Challenges in MU datasets are Variation in scale, rotation, illumination, shape of hands and similarities between gesture poses like a) pose 'm', b) pose 'n'.

TABLE I
Performance comparison of different features extracted from pre-trained AlexNet with proposed technique on MU dataset

Pre-trained CNN feature layer	LOO CV	Holdout CV
AlexNet 'FC7'	80.87 ± 2.11	98.73 ± 1.92
AlexNet 'FC6'	85.02 ± 1.75	98.97 ± 1.60
AlexNet 'FC7'+'FC6'	84.51 ± 1.70	99.15 ± 1.87
Proposed (AlexNet 'FC6'+ PCA)	87.83 ± 1.79	99.32 ± 1.40

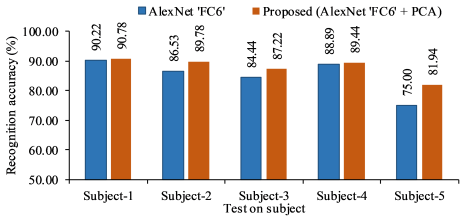


Fig. 5. Subject wise recognition accuracy of LOO CV test on the benchmark dataset.

IV. Experimental results

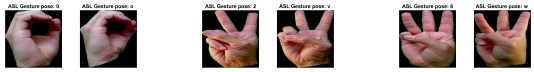
All experiments were conducted on an Intel Xeon 2.40-GHz CPU with 24GB RAM and 8GB NVIDIA GPU card. The performance of proposed technique is evaluated in terms of mean accuracy and it is compared with earlier reported techniques.

A. Performance Evaluation

Table I shows the performance comparison of features extracted from FC layers of pre-trained AlexNet with the proposed technique using both the CV test on MU dataset. The tabulation result shows that, the deep features extracted from AlexNet 'FC6' is superior than 'FC7' and also the combination of both feature 'FC7' + 'FC6' in LOO CV test. Since the holdout CV test is a biased one hence, the performance in all above features are nearly equal. PCA based dimension reduction technique is applied on AlexNet 'FC6' feature to reduced the redundant in the feature vector. In this analysis the variance of sum of deep features above 99.9% are considered as redundant feature. The final reduced feature is used to find the gesture recognition performance. The tabulation result shows that the performance of the proposed technique is 87.83% mean accuracy which is superior than the mean accuracy result of Alexnet 'FC6' feature. The subject wise recognition performance is shown in Fig. 5. The recognition accuracy of all subjects using the proposed technique shows superior than without feature dimension reduction technique. Therefore, it proves that the recognition accuracy improves by the application of dimension reduction technique of the extracted deep feature.

B. Error analysis

In the ASL data, there are many similarities between the ASL digit and alphabet sign language representation. Therefore, the most confuse gesture poses in the proposed result are '2' and 'v', '0' and 'o', '6' and 'w' are shown in Fig 6.



(a) (b) (c)

Fig. 6. The Most confusing ASL gesture poses in the proposed work of the dataset are (a) '0' and 'o', (b) '2', and 'v', (c) '6', and 'w'.

TABLE II
COMPARISON OF ALEXNET 'FC6' FEATURE WITH PROPOSED TECHNIQUE USING LOO CV TEST ON TWO SEPERATE MODELS SUCH AS ASL DIGITS AND ALPHABETS

ASL gesture pose	Number of classes	AlexNet 'FC6'	Proposed (AlexNet 'FC6' + PCA)
ASL digit	10	93.26 ± 1.08	95.00 ± 1.20
ASL alphabets	26	90.82 ± 1.22	92.60 ± 1.17

From the figure it is difficult to distinguish the gesture poses for a person in human eye due to similarities in the gesture representation.

Therefore in this work, two different models are developed to recognized the ASL gesture poses, one for digit and other for alphabet recognition. The gesture recognition performance of two models using the proposed technique is presented in Table II. The result shows that the proposed technique shows the superior result than AlexNet 'FC6' deep feature. Hence, it is better to use two separate models for ASL gesture recognition.

C. Comparison with earlier methods

The performance of proposed technique is compared with earlier reported technique using LOO CV and holdout CV test and the results are presented in Table III and Table IV respectively. As shown in the Table III, the recognition accuracy of proposed technique is 13.97% and 3.81% higher in mean accuracy than CNN amd FCNN technique respectively. Similarly, the mean accuracy performance of proposed technique in holdout CV test shows superior performance than eralier reported technique as shown in Table IV.

V. CONCLUSION

In this work, a PCA based reduced deep CNN feature is proposed for recognition of static hand gesture images. The deep features are extracted from fully connected layers of pre-trained AlexNet. Then PCA dimension reduction technique is used to reduce the redundant features in the feature vector. The proposed technique does not required any hand segmentation or localization technique. The experiments are conducted on 36 ASL gesture poses using LOO CV and holdout CV test. The experimental result shows that the performance of the proposed technique (AlexNet 'FC6' + PCA) shows superior than the features extracted from individual FC-layer 'FC6', 'FC7' and also earlier reported techniques. it is also observed that, the gesture recognition performance of proposed technique for two separate ASL digit and alphabets classes shows higher mean accuracy in LOO CV test.

TABLE III
COMPARISON OF EARLIER TECHNIQUES WITH PROPOSED TECHNIQUE USING LOO CV TEST

Test methods	Mean Accuracy (%)
CNN [8]	73.86 ± 1.04
FFCN [8]	84.02 ± 0.59
Proposed	87.83 ± 1.79

TABLE IV
COMPARISON OF EARLIER TECHNIQUES WITH PROPOSED TECHNIQUE USING HOLDOUT CV TEST

Test methods	Mean Accuracy (%)
HU [2]	37.02
ZM [2]	91.08
GB [2]	97.15
GB-ZM [2]	97.09
GB-HU [2]	97.63
Proposed	99.32

ACKNOWLEDGMENT

The authors would like to thank the Defence Research and Development Organisation (DRDO) India for providing financial support during the course of this research project (Letter No. ERIP/ER/13006034/M/01/1609).

REFERENCES

[1] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152–165, Dec. 2015.

[2] S. F. Chevtchenko, R. F. Vale, and V. Macario, "Multi-objective optimization for hand posture recognition," *Expert Systems with Applications*, vol. 92, pp. 170–181, Feb. 2018.

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

[4] J. P. Sahoo, S. Ari, and D. K. Ghosh, "Hand gesture recognition using DWT and F-ratio based feature descriptor," *IET Image Processing*, vol. 12, no. 10, pp. 1780–1787, Oct. 2018.

[5] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Sep. 2012.

[6] D. L. Lee and W. S. You, "Recognition of complex static hand gestures by using the wristband-based contour features," *IET Image Processing*, vol. 12, no. 1, pp. 80–87, Jan. 2018.

[7] A. Chaudhary and J. Raheja, "Light invariant real-time robust hand gesture recognition," *Optik*, vol. 159, pp. 283–294, Apr. 2018.

[8] 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.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] S. Kundu and S. Ari, "P300 detection with brain–computer interface application using PCA and ensemble of weighted SVMs," *IETE Journal of Research*, vol. 64, no. 3, pp. 406–414, Aug. 2017.

[11] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[12] A. Barczak, N. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2D static hand gesture colour image dataset for ASL gestures," *Research Letters in the Information and Mathematical Sciences*, vol. 15, pp. 12–20, 2011.