Deep Feature Extraction and Multi-feature Fusion for Similar Hand Gesture Recognition

Cunhuang Xie
School of Electron. Inf. & Commun.,
Huazhong Univ. of Sci. & Tech.,
Wuhan, Hubei, China 430074
xch382@hust.edu.cn

Li Yu
School of Electron. Inf. & Commun.,
Huazhong Univ. of Sci. & Tech.,
Wuhan, Hubei, China 430074
hustlyu@hust.edu.cn

Shengwei Wang
School of Electron. Inf. & Commun.,
Huazhong Univ. of Sci. & Tech.,
Wuhan, Hubei, China 430074
kadinwang@hust.edu.cn

***Abstract*—Gesture recognition plays an important role in human computer interaction, but the accuracy is unsatisfactory in complex gestures with slight discrimination. In this paper, a framework facing to recognize complex and similar gestures is presented. In the framework, a parallel connection structure of convolutional neural network (CNN) is designed to extract deep features of complex and similar gestures from RGBD images. Then, a novel feature fusion method is proposed to achieve multi-feature fusion and dimension reduction simultaneously. According to experimental results on American Sign Language (ASL) dataset, the proposed framework reaches 99.042% recognition rate and outperforms current state-of-the-art methods.**

***Index Terms*—Feature fusion, gesture recognition, locally linear embedding, parallel connection structure**

I. Introduction

Hand gesture recognition is one of the most active fields in computer vision. Generally, the system of hand gesture recognition can fall into three categories: sensor-based methods, vision-based methods and hybrid methods. For hybrid systems, the visible cameras and other kinds of information acquisition devices, such as depth sensors, are used to extract multimodal features. Gesture recognition based on 2D images usually requires a clean background, which limits its real-world applications. The hybrid methods can overcome this problem by utilizing multimodal information such as depth cues. Under the circumstance, more research work in the field of recognition proposes to combine color images and the corresponding depth information.

Isaac *et al.* [1] proposed a system for American Sign Language(ASL) [2] finger-spelling recognition. The system extracted the wavelet features of RGBD hand images and feeds them into neural networks for classification. Nicolas *et al.* [2] proposed an interactive hand recognition system including hand tracking and gesture recognition. Hand shape features from color images and depth data were fed into random forest to obtain classification result. Estrela *et al.* [3] combined intensity and shape information based on the bag of features with the partial least squares. Then, models for each letter in the ASL dataset were created for classification. Zhang *et al.* [4] proposed a characteristic descriptor named Histogram of 3D Facets (H3DF) to explicitly encode the 3D shape information from the depth data. Apparently, the above researches focus more on the application and improvement of handcrafted descriptors, which can't provide abundant deep level information. Besides, they consider RGB images and depth data independently and utilize multimodal features without a reasonable strategy.

In this paper, we present a novel recognition framework for discriminative feature extraction. To extract the multimodal features, we design a parallel connection structure of convolutional neural network (CNN). The special structure of the deep-feature learning model offers a new way to extract unified-paradigm features from RGB and depth channels. A novel feature fusion method is then used to combine the deep features with other local descriptors. The feature fusion method can handle different types of features and reduce the dimension of feature vectors. Finally, a SoftMax classifier for similar gesture classification is trained to measure the performance. Experimental results show that the proposed framework is reasonable and effective.

The remainder of this paper is organized as follows. Section II describes the detailed procedures of our proposed method. Experimental results and discussion are shown in Section III. Finally, the conclusion is drawn in Section IV.

II. Proposed Method

The proposed framework of gesture recognition is outlined in Fig. 1. The discriminative deep features are firstly extracted through a parallel connection structure of convolution neural network, which is described in Section II-A. Then, a novel feature fusion method is proposed to achieve multiple features fusion and dimension reduction, which is given in Section II-B.

A. Discriminative Deep-feature Learning Model

In this paper, we consider that both depth and RGB channels are equally important to perform sign language recognition and can be task-dependent. Based on this perspective, we design a parallel connection structure to extract discriminative deep features for complex and similar gestures.

The network for RGBD images consists of two sub-networks in the first five layers, as shown in Fig. 1. Then, features learned from two sub-networks are concatenated to form the input feature vector of the coupling layer. Namely, the network is trained as a whole from the coupling section. To

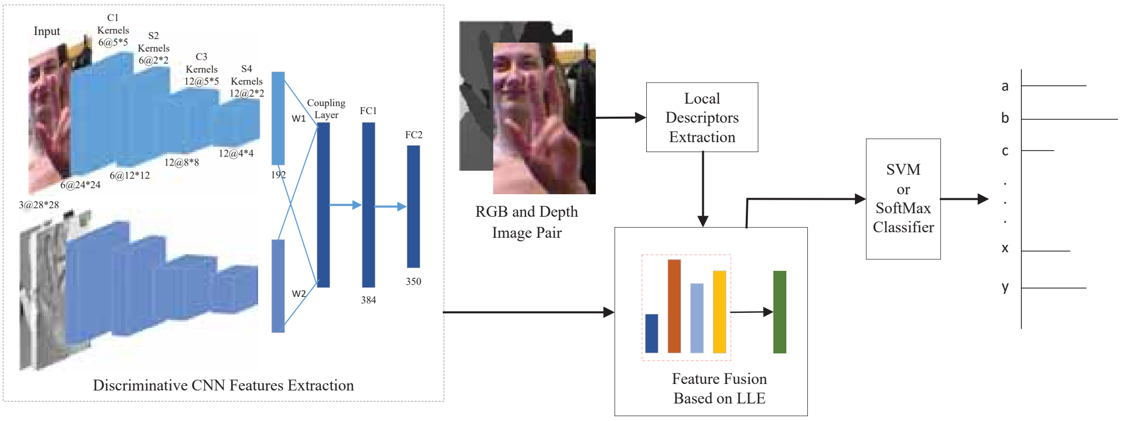

Fig. 1. Pipeline of proposed human gesture recognition framework.

measure the different importance of two sub-networks, w_1 and w_2 are introduced to represent the weights of color and depth data, respectively. These two parameters can be regarded as the contribution of two sub-networks to the final combinational deep feature representation. Apparently, w_1 and w_2 can be updated by stochastic gradient descent (SGD). Benefiting from the special parallel connection structure, we can extract a potential combinational deep feature representation from RGBD images in a real sense.

As for the input of sub-network of depth channel, the HHA [5] is proposed to encode depth images. Fig. 2 shows the RGB and depth image pair from ASL dataset and corresponding HHA images.

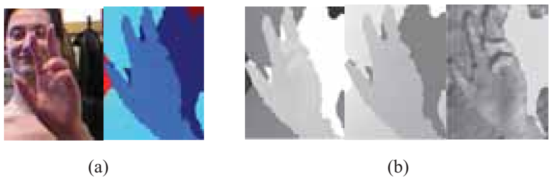

Fig. 2. (a) RGB and depth image pair. (b) Corresponding HHA encoded depth images with three channels.

B. Feature Fusion Method

Deep learning methods require large-scale datasets and these datasets might not be easy to obtain. Besides, local information is important for similar gesture recognition. However, the information may be significantly eliminated in the highly compressed representation at the top of fully-connected (FC) layer [6]. Fortunately, local feature descriptors contain complete local information and they may be complimentary to deeply learned features [7].

In this paper, a novel feature fusion method is proposed to fuse the deep features with some local feature descriptors. Since deep features are similar in appearance to image gradient, we choose histogram of oriented gradients (HOG) [8] as one of local descriptors to be fused. Histogram of oriented normal vectors (HONV) [9] is also chosen for depth map feature extraction. Since HOG feature has no rotation-invariant characteristic and the classes of gestures should not be changed with random angle rotation, we extract HOG feature in a group of concentric sector region. The feature improved from HOG is called HOG based on sector region (HOGS) and it is also chosen for fusing.

The proposed feature fusion method is improved based on locally linear embedding (LLE) [10]. Compared to LLE which only applies to one feature space, the proposed method can handle multimodal information from different feature space with the characteristic of dimension reduction. Traditional LLE maps the data set $X = \{x_1, x_2, \dots x_n\} \in \mathbb{R}^{d \times n}$ to a data set $Y = \{y_1, y_2, \dots y_n\} \in \mathbb{R}^{m \times n}$. LLE firstly constructs each sample x based on its K nearest neighbors and then obtain the reconstruction weight matrix W. The cost function to optimize can be described as

$$\varepsilon(X) = \sum_{i=1}^{n} \left| x_i - \sum_{j=1}^{K} w_{ij} x_{N(j)} \right|^2, \tag{1}$$

where $x_{N(j)}$ is the jth neighbor of x_i which will be reconstructed based on its neighbors $x_{N(1)}, x_{N(2)}, \dots, x_{N(K)}$. w_{ij} are computed using the square principle and the sum of them is 1. If x_i and x_j are not neighborhoods, $w_{ij} = 0$.

Once the weight matrix W is fixed, the new dataset Y can be sought by solving the following minimization problem

$$\begin{aligned} \min \varepsilon(Y) = \min \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{K} w_{ij} y_{N(j)} \right|^2 &= tr(YMY^T) \\ \text{subject to} \qquad \tfrac{1}{n} YY^T = I, & \end{aligned} \tag{2}$$

where $M = (I - W)^T (I - W)$ and y_i is the columns of Y. Apparently, it is an eigenvalue problem. All eigenvectors of the matrix M are the solution to Eq. (2).

In proposed feature fusion stage, we want to fuse four features including deep feature and three local feature descriptors. According to Eq. (1), four weight matrices $\{W^c, c = 1, 2, ..., 4\}$ can be acquired on the training dataset. Weight factors $\{w_c, c = 1, 2, ..., 4\}$ are then assigned to four weight matrices. Namely, various feature representations have different importance. The mapping dataset $Y = \{y_1, y_2, \ldots y_n\} \in \mathbb{R}^{m \times n}$ can be sought by the following optimization problem

$$\begin{aligned} &\min \varepsilon(Y) = tr\{Y(I - \sum_{c=1}^{4} w_c W^c)^T (I - \sum_{c=1}^{4} w_c W^c) Y^T\} \\ &\text{subject to} \quad \tfrac{1}{n} Y Y^T = I, \quad 0 \le w_c \le 1, \quad \sum_{c=1}^{4} w_c = 1. \end{aligned} \tag{3}$$

When applying the fusion algorithm to classification, it requires the storage of whole training set to construct the weight matrix. To solve this problem, a transformation is adopted as

$$Y = F^T X. \tag{4}$$

Moreover, we hope that the transformation can help to narrow the intra-class distance and enlarge the inter-class distance, which can be described as

$$y_{ij}(\tau - d_{ij}(F)) > \varepsilon, \tag{5}$$

Where $d_{ij}(F)$ is the relative feature distance between two gesture samples y_i and y_j, τ is the average distance between samples of the same class and ε is a threshold parameter. if y_i and y_j are from the same class ($y_{ij} = 1$), their distance should be smaller than $\tau - \varepsilon$, otherwise ($y_{ij} = -1$) the distance should be larger than $\tau + \varepsilon$. $d_{ij}(F)$ can be described as

$$\begin{aligned} d_{ij}(F) &= \|y_i - y_j\|_2^2 \\ &= (F^T(x_i - x_j))^T (F^T(x_i - x_j)). \end{aligned} \tag{6}$$

Therefore, a discriminative term can be introduced to the minimization problem in Eq. (3) to narrow the intra-class distance and enlarge the inter-class distance, which can be defined as

$$D(F) = \sum_{\forall i,j} \max(0, \varepsilon - y_{ij}(\tau - d_{ij}(F))). \tag{7}$$

Minimization problem in Eq. (3) can be rewritten as

$$\begin{aligned} \min \varepsilon(Y) \;&= tr(F^T X(I - \sum_{c=1}^{4} w_c W^c)^T (I - \sum_{c=1}^{4} w_c W^c) X^T F) \\ &\quad + \alpha \sum_{\forall i,j} \max(0, \varepsilon - y_{ij}(\tau - d_{ij}(F))) \\ &s.t. \;\; \tfrac{1}{n} F^T X X^T F = I, \;\; 0 \le w_c \le 1, \;\; \sum_{c=1}^{4} w_c = 1. \end{aligned} \tag{8}$$

Eq. (8) is an optimization problem with two variables including F and $\{w_c, c = 1, 2, ..., 4\}$, which is difficult to obtain the global optimal solutions directly. But by fixing F, the optimization problem can be transformed as

$$\begin{aligned} \min \varepsilon(Y) \;&= tr(F^T X(I - \sum_{c=1}^{4} w_c W^c)^T (I - \sum_{c=1}^{4} w_c W^c) X^T F) \\ &s.t. \qquad 0 \le w_c \le 1, \quad \sum_{c=1}^{4} w_c = 1, \end{aligned} \tag{9}$$

when $\{w_c, c = 1, 2, ..., 4\}$ is fixed, the problem can be transformed as

$$\begin{aligned} \min \varepsilon(Y) \;&= tr(F^T B F) + \alpha \sum_{\forall i,j} \max(0, \varepsilon - y_{ij}(\tau - d_{ij}(F))) \\ &s.t. \quad \tfrac{1}{n} F^T X X^T F = I, \end{aligned} \tag{10}$$

Where $B = X(I - \sum_{c=1}^{4} w_c W^c)^T (I - \sum_{c=1}^{4} w_c W^c) X^T$ is a constant. Eq. (10) can be transformed into an unconstrained optimization problem which can be described as

$$\begin{aligned} \min \varepsilon(Y) \;&= tr(F^T B F) + \alpha \sum_{\forall i,j} \max(0, \varepsilon - y_{ij}(\tau - d_{ij}(F))) \\ &\quad + \tfrac{1}{2} \left\| F^T X X^T F - nI \right\|_2^2. \end{aligned} \tag{11}$$

Eq. (9) can be solved by quadric programming and Eq. (11) can be optimized by SGD. So the matrix F can be updated based on the cross-iteration optimization between Eq. (9) and Eq. (11). Then the training and testing sets are transformed by Eq. (4). Finally, we can train the classifier for similar gesture recognition based on the mapped datasets.

III. Experimental Results

A. Datasets

In this paper, we choose ASL dataset. ASL is a finger-spelling dataset where each letter of alphabet is represented by an unique hand gesture. As shown in Fig. 3, the difference of these gestures is only the position of fingers so that they are extremely similar visually.

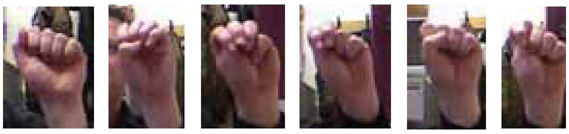

Fig. 3. Illustration of some similar gestures from ASL.

The dataset contains 24 categories except for letters j and z which are built with multiple frames. All gestures are captured by 5 subjects with Microsoft Kinect sensors. Each gesture contains about 1000 samples including color images and corresponding depth maps for each subject. 80% of the samples are used for training and other 20% for testing. 20% of the training set are reserved to construct validation set.

B. Extraction of Deep Features

In this paper, the structure of CNN designed for the extraction of deep features is shown in Fig. 1. The input color images and the corresponding HHA encoded depth images are resized to 28×28 and the extracted features are 350 dimensional vectors. For the parameter settings of SGD, The momentum coefficient is 0.9. Learning rate is initially 0.1 and divided by a factor of 2 if the training error stops making progress for 10 epochs on the validation set.

Stochastic pooling is chosen for a better performance. To avoid overfitting and improve the generalization of our model, we apply the dropout to the output of the fully connected layers and the value of dropout rate is 0.5.

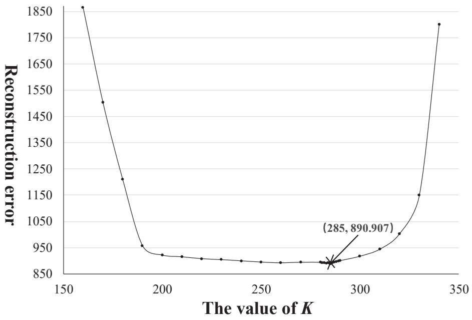

Fig. 4. Reconstruction error with different value of K.

C. Results of Feature Fusion and Comparison

In the experiments, we choose $K = 285$ that can minimize the average reconstruction error of four feature spaces based on Eq. (1). The Fig. 4 shows the reconstruction error when choosing different value of K. The dimension of fused feature is set to 2055 which is computed in a way the similar to PCA. Table I summarizes the classification results of different features including the fused feature.

TABLE I
RECOGNITION ACCURACIES OF DIFFERENT FEATURES.

Method	Training set(%)	Testing set(%)
HOG	100	92.083
HONV	100	89.875
HOGS	100	91.250
CNN	100	97.917
Fused	100	99.042

Table I shows that the performances of first three descriptors are similar and the CNN deep feature is much better than those manually designed. The recognition accuracy of fused feature outperforms all the others since these local descriptors can be seen as the complement of the deep feature and further improve its ability of classification.

We compare our performance with various methods proposed on the ASL dataset including HSF+RDF [2], SIFT+PLS [3], DD+RDF [11] and H3DF+SVM [4]. HSF refers to hand shape feature, RDF refers to random decision forest, PLS refers to partial least square, DD refers to depth difference and H3DF refers to the histogram of 3D facet. The results are shown in Fig. 5.

It can be found that our proposed framework behaves better than four state-of-the-art approaches, which reaches 99.042% recognition rate. Different from the other four methods, our proposed framework adopts deep feature learning model which is powerful for classification. Considering the importance of local information, we proposes a novel feature fusion method. Three local descriptors are then applied for fusion. The fusion method can make the deep feature more discriminative. When related features have similar classification performances, the fusion method can also reduce redundant information by dimension reduction. The achieved classification performances demonstrate the proposed framework successfully improves the gesture recognition rate.

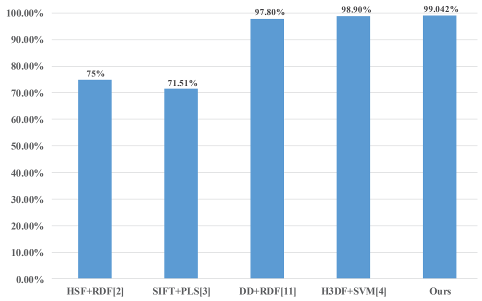

Fig. 5. Recognition accuracy of different methods on the ASL dataset.

IV. Conclusion

In this paper, we presented a framework for hand gesture recognition. Our proposed method considered the discriminative feature extraction for gestures which were extremely similar visually. To extract a deep feature representation of the complex gestures, a parallel connection structure of CNN was designed for utilizing RGBD images. A novel multi-feature fusion algorithm was then proposed for multimodal feature extraction, which also achieved the dimension reduction. The experimental results showed that the proposed method outperformed other state-of-the-arts on the ASL finger-spelling dataset.

V. Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 61871437).

References

[1] J. Isaacs and S. Foo, “Hand pose estimation for american sign language recognition,” in *Proc. IEEE Thirty-Sixth Southeastern Symp. Syst. Theory*, 2004, pp. 132–136.

[2] N. Pugeault and R. Bowden, “Spelling it out: real-time asl fingerspelling recognition,” in *IEEE Int. Conf. Comput. Vision Workshops*, 2011, pp. 1114–1119.

[3] B. Estrela, G. Cámara-Chávez, M. F. Campos, W. R. Schwartz, and E. R. Nascimento, “Sign language recognition using partial least squares and RGB-D information,” in *Proc. IX Workshop de Visao Computacional, WVC*, 2013.

[4] C. Zhang and Y. Tian, “Histogram of 3D facets: a depth descriptor for human action and hand gesture recognition,” *Comput. Vision and Image Understanding*, vol. 139, pp. 29–39, 2015.

[5] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, “Learning rich features from RGB-D images for object detection and segmentation,” in *European Conf. Comput. Vision*, 2014, pp. 345–360.

[6] S. Guo, W. Huang, L. Wang, and Y. Qiao, “Locally supervised deep hybrid model for scene recognition,” *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 808–820, 2017.

[7] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[9] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, “Histogram of oriented normal vectors for object recognition with a depth sensor,” in *Asian conf. comput. vision*, 2012, pp. 525–538.

[10] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Sci.*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[11] C. Keskin, F. Kiraç, Y. E. Kara, and L. Akarun, “Randomized decision forests for static and dynamic hand shape classification,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition Workshops*, 2012, pp. 31–36.