Multimodal Gesture Recognition Based on the ResC3D Network

Qiguang Miao[1] * Yunan Li[1] Wanli Ouyang[2] Zhenxin Ma[1] Xin Xu[1] Weikang Shi[1] Xiaochun Cao[3]
[1] School of Computer Science and Technology, Xidian Univeristy
[2] Department of Electronic Engineering, The Chinese University of Hong Kong
[3] State Key Laboratory of Information Security,
Institute of Information Engineering, Chinese Academy of Sciences

Abstract

Gesture recognition is an important issue in computer vision. Recognizing gestures with videos remains a challenging task due to the barriers of gesture-irrelevant factors. In this paper, we propose a multimodal gesture recognition method based on a ResC3D network. One key idea is to find a compact and effective representation of video sequences. Therefore, the video enhancement techniques, such as Retinex and median filter are applied to eliminate the illumination variation and noise in the input video, and a weighted frame unification strategy is utilized to sample key frames. Upon these representations, a ResC3D network, which leverages the advantages of both residual and C3D model, is developed to extract features, together with a canonical correlation analysis based fusion scheme for blending features. The performance of our method is evaluated in the Chalearn LAP isolated gesture recognition challenge. It reaches 67.71% accuracy and ranks the 1st place in this challenge.

1. Introduction

Gesture recognition has been a promising topic since it has many applications, such as visual surveillance, video retrieval and human-computer interaction (HCI). In gesture recognition, the main task is to extract features from an image or a video and to issue a corresponding label.

Although in the past decades, many methods have been proposed for this issue, ranging from static to dynamic gestures, and from motion silhouettes-based to the convolutional neural network-based, there are still many challenges associated with the recognition accuracy. The gesture-irrelevant factors, such as the illumination, the background, the skin color and clothes of performers can handicap the recognition of gestures. Furthermore, it can be more arduous when the task is recognizing dynamic gestures in videos. The velocity and angle of performers showing a gesture can be different since there are no standards for the gesture performing. The increasing number of classes can also be a difficulty for the overlap between classes will be higher. A good gesture recognition approach should be able to generalize over intra-class variations and distinguish inter-class ones. Therefore, extracting discriminative spatiotemporal features plays a crucial role in accomplishing the recognizing task.

In this paper, we propose a multimodal gesture recognition method using a ResC3D network [32] for large-scale video-based gesture recognition. We first perform a video enhancement on both RGB and depth data (which is captured concurrently with the RGB counterpart by Kinect) to normalize the illumination and denoise. Then we propose a scheme of weighted frame unification to sample the most representative frames for identifying gesture. That scheme is based on key frame attention mechanism, which deems the movement intensity as an indicator for selecting frames. Then multimodal data, including RGB, depth and optical flow data generated from the RGB ones, are sent to the ResC3D network, which is based on the work of Tran *et al*.[31] and He *et al*.[8], to extract spatiotemporal features. Finally, the features are blended together in terms of a statistical analysis method - canonical correlation analysis, and the final recognition result is obtained by a linear SVM classifier. The pipeline of our method is depicted in Figure 1. Our main contributions can be summarized as below:

- A pre-processing of video enhancement. Since the RGB videos are captured under different environments, the illumination condition is a gesture irrelevant variable. Meanwhile, the depth videos also suffer from noise. Therefore, we first employ Retinex and median filter for RGB and depth videos to eliminate the influence of illumination variation and noise.

- A weighted frame unification scheme. Convolutional neural networks require a fixed input dimension, thus we need to unify the frame numbers before sending

*Corresponding author

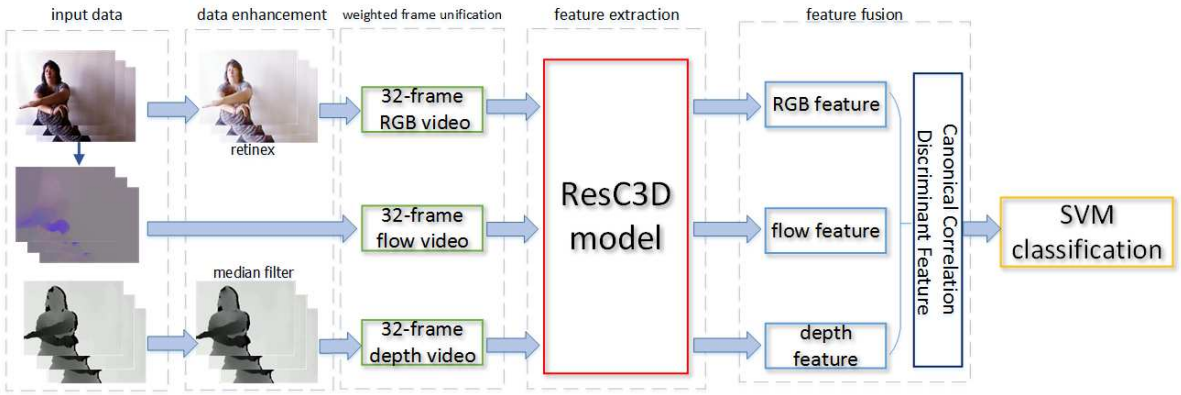


Figure 1. The pipeline of our method.

them to the network. However, it is crucial to preserve the motion information while sampling the videos. Based on an observation that we call "key frame attention mechanism", we utilize the optical flow to characterize the intensity of movement, and weight different parts of the video when select frames.

- A ResC3D model for learning and feature extracting. We employ a ResC3D model to learn and extract features of various data, which leverages the benefits of ResNet [8] and C3D [31].

- A statistical analysis based fusion scheme. We adopt a canonical correlation analysis based method, which analyzes the pair-wise correlation between features from different modalities to fuse the features together.

2. Related Works

Gesture taxonomies and representations have been studied for decades. The vision based gesture recognition techniques include the static gesture oriented and the dynamic gesture oriented methods [26]. To recognize static gestures, namely postures in still images, a general classifier like random forest [28] or template-matching method [25] is enough. As the dynamic gesture recognition has a temporal aspect, more endeavors should be made to demonstrate the motion in videos. In the early stage, the silhouette of performers is used for global gesture recognition. Two famous indicators are motion energy image (MEI) and motion history image (MHI) [1], which represents where motion occurs or a recency function of the silhouette motion, respectively. As an extension, Weinland *et al*.[40] combine the silhouettes taken by multiple cameras to construct motion history volumes. Some of the handcrafted features, such as histogram of oriented gradient (HOG) and histogram of optical flow (HOF) are thereafter employed to describe gestures [17, 21, 22]. Then some researchers extend those handcrafted features into spatiotemporal ones to handle videos more effectively. Klaser *et al*.[16] propose a 3D HOG feature for action recognition. Sanin *et al*.[27] use a spatiotemporal covariance descriptor to achieve gesture recognition. Wan *et al*.extend scale-invariant feature transform (SIFT) to 3D enhanced motion SIFT [35] and propose the mixed features around sparse keypoints (MFSK) for one-shot learning in gesture recognition [33]. Meanwhile, techniques that handle the addition temporal dimension like hidden Markov models (HMMs) [22, 41], condition random field (CRF) [39] and dynamic time warping (DTW) [3] are applied to model gestures. Automata-based methods are another alternative to solve this issue in the literature. Finite state machines (FSMs) [44, 9] is commonly employed, in which the states can represent the postures whereas the transitions are used to represent the motion information.

Another way to handle gesture recognition is learning based algorithms. Convolutional neural networks (CNNs) can build high-level features from low-level images, and is invariant to rigid transformation, therefore many gesture recognition tasks are based on it. Nagi *et al*.[24] use max-pooling convolutional neural networks for real-time hand gesture recognition in human-robot interaction. Karpathy *et al*.[15] classify videos on a large-scale dataset with a CNN-based model. Simonyan and Zisserman [29] propose a two-stream network to extract spatial concurrent with temporal features. On the basis of their work, Wang *et al*.[36] develop a temporal segment network which exploits RGB and optical flow data for extracting spatial and temporal features. As the traditional CNN can only deal with 2D images instead of videos, researchers try to modify the structure of convolutional layer to handle temporal features. Ji *et al*.[12] employ a hardwired layer to extract manual features like optical flow and gradient and use 3D CNN to learn the features further. Tran *et al*.[31] propose a more interesting 3D CNN model, called C3D, which is based on BVLC caffe [13]. Their method can process on videos directly and show a great promise in action or gesture recognition tasks even on

large-scale datasets. In [32], they combine it with ResNet and that network seems more efficient. There are many methods [19, 42, 5] developed based on the C3D model. Since the input is sequence data in the video-based recognition, Recurrent neural networks (RNN) are also employed in gesture recognition. Molchanov *et al*.[23] combine C3D and RNN to form a recurrent 3D CNN for video-based detection and classification issues. Donahue *et al*.[4] extract traditional CNN features first and then utilize a LSTM network for video labeling or caption. Zhu *et al*.[46] combine C3D and LSTM for learning features.

3. Gesture recognition with ResC3D model

As depicted in Figure 1, the overall process can be divided into five parts: video enhancement, weighted frame unification scheme, multimodal feature extraction with ResC3D model, canonical correlation analysis based feature fusion, and finally a SVM classifier. The optical flow videos are first generated from the RGB data stream in terms of Brox *et al*.[2]. The flow data is used as another modality of data that concerns about the motion path. Then a video enhancement pre-processing is manipulated on both RGB and depth data. Such a process is proposed to eliminate the variance of illumination of RGB data and denoise for the depth one. After that, the number of input frames is unified since the fixed dimension is required by convolutional networks. However, in order to preserve the motion information as much as possible, we propose a weighted frame unification scheme to select "key frames" according to the movement conditions. Following, a ResC3D model, which adopt the strengths of deep residual network and C3D model, is utilized to extract spatiotemporal features for all the input videos. As our method is based on multimodal inputs, we need to make a comprehensive prediction based on all the data. Therefore, we use canonical correlation analysis to maximize the pair-wise correlation among features from different modalities and blend them together. At last, a linear SVM classifier is used to give the final classification result.

3.1. Video enhancement

As mentioned before, there are many variants that have no relation with gestures in the RGB videos. The illumination is one sort of it. As shown in Figure 2, for videos that share the same label, once they are captured in different places, the illumination may vary a lot. The dusky environment even makes some videos hard to be recognized. Undoubtedly, that is a huge barrier for gesture recognition. We turn to Retinex theory for illumination normalization. Retinex theory [18] is a form of color constancy, which indicates the perceived color of objects remains relatively constant under varying illumination conditions. According to that theory, the observed illumination of an object is determined by the reflectance light from the object surface and environmental illumination:

Figure 2. An example of our video enhancement on RGB (left part) and depth (right part) data. The first row represents videos without pre-processing. It is obvious that the illumination of RGB videos with different performers can vary a lot. The depth videos mainly suffer from noises along with edges. Retinex is processed on RGB and median filter on depth videos. The result on the second row shows our strategy is effective to enhance both videos.

$$I(x) = L(x) \times R(x) \tag{1}$$

where x indicates a position in the image, $I(x)$ is the observed intensity of the image, $L(x)$ and $R(x)$ represent the intensities of object reflectance and environmental illumination, respectively. The enhancement can be achieved by eliminating $L(x)$ and recovering $R(x)$ as:

$$R(x) = exp(log(I(x)) - log(L(x))) \tag{2}$$

As $L(x)$ is hard to obtain directly from a single image, it is always approximated by filtering $I(x)$ with low pass filter [10]. In this paper, three scales of Gaussian filter with $\sigma = 15$, $\sigma = 80$, and $\sigma = 250$ are employed for obtaining it.

Compared with the RGB counterparts, the noise is a major problem for depth videos as illustrated in the right part of Figure 2, especially on the edges marked in the red box. That is because the Kinect cannot estimate very well on the depth discontinuous regions, and the noise (usually performs as black points) are apparent in these regions.

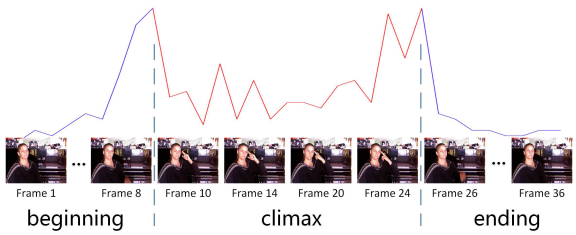


Figure 3. An example of a RGB gesture video. The motion in the beginning and end stage of the video is slight. However, in the intermediate 18 frames, the movement is significant. That is a crucial part for recognizing a gesture.

To eliminate the noise while preserving the edges, we employ a common but fast to implement filter in digital image processing - median filter. The result of median filter

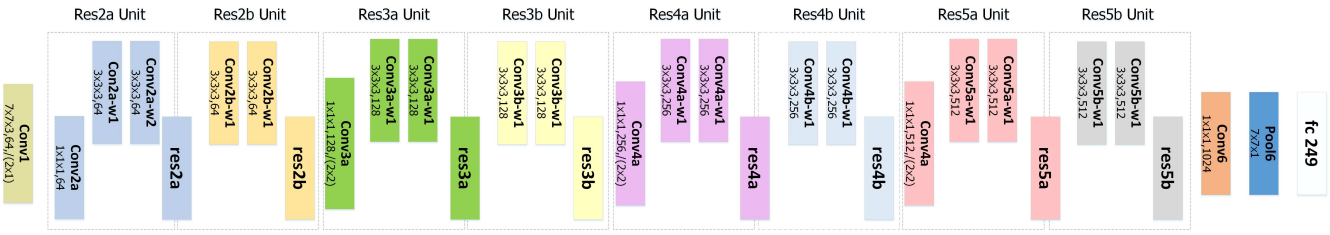


Figure 4. Structure of ResC3D network. It consists of 8 residual units that correspond to 8 convolutional layers in the C3D model [31]. Following a linear projection is achieved by a $1 \times 1 \times 1$ convolutional layer for dimension expansion. The network ends with a spatial global average pooling layer.

is shown below the original depth data in Figure 2. With the help of that filter, it seems to be more smooth in these regions and the influence of noise is alleviated a lot.

3.2. Weighted frame unification

Convolutional networks require fixed dimension inputs, which means the length of videos should be uniform. Therefore, for long videos, the information of the motion path is inevitably lost. To preserve the motion information as much as possible with limited frame numbers is a challenging task, and that is what we aim to solve in this subsection.

Noticing that in most videos with a single gesture, the action can be intuitively generalized as three phases - beginning, climaxing and end. Take the gesture in Figure 3 as an example, between frame 1 and frame 8, the performer's movement is slight while preparing for the gesture. Then in frame 9 and 10, she raises her arm and becomes ready for the posture. The motion is continued for performing the target gesture until the frame 26, during which the movement is relatively drastic. After that, the performer's arm lays back and the whole gesture is over. We can learn that the frames in different phases share different importance. In the beginning and end, the movement for preparing or returning stage is slight, and it is of less importance to distinguish this gesture, whereas that in the climax stage is sharp and crucial to identify the gesture. Consequently, the frames of the climax stage are key frames to the gesture and we should pay more attention to them. We call that "key frame attention" mechanism. Based on that, we develop a weighted frame selection method. For a given video S, we first cut it into n sections, and calculate the average optical flow of each section S_1, S_2,..., S_n, since the average optical flow is leveraged as an indicator of movement. As the optical flow videos is obtained following Brox *et al.*[2], the average optical flow can be calculated frame-by-frame and be added up for each section. Then the importance of section i $(1 \leq i \leq n)$ is calculated as the ratio of this section's average value of the optical flow and the sum of that for the whole video. That is also the proportion of the frame amount for this section in the frame number unified video. In this paper, we set the unification frame number (i.e., the normalized length of videos for the network) as 32, and n as the rounding up ratio of frame numbers of the original video and half of the unification benchmark, namely 16 in this paper.

3.3. Learning and feature extraction based on ResC3D model

As the data is acquired, we can train a model to extract features for classification. In recent years, the C3D model [31] has been proved efficient in video-based recognition tasks. Meanwhile, He *et al.*'s residual network [8] also shows great promise to solve the degradation problem with training deep networks. In this paper, we utilize a model that combines these two networks, namely ResC3D model, which can help to boost the performance. The structure of ResC3D is depicted in Figure 4. Most of our convolutional layers are with $3 \times 3 \times 3$ filters, which process on the spatial as well as temporal domain. Similar as [31], the number of filters are set to 64, 64, 64, 128, 128, 256, 256, 512, 512, and 1024 for an additional layer with $1 \times 1 \times 1$ kernel size to project the feature into a higher dimension. The pooling layer is also replaced by convolutional layers with a stride of 2 to achieve downsampling at conv3a, conv4a, and conv5a. Then a spatial global average pooling layer with kernel size 7×7 is performed. Finally, a fully-connected layer is used to corresponding to 249 classes.

3.4. Fusion scheme

The fusion of information is important for multimodal gesture recognition, which can occur on either data level, feature level or decision level. However, the data level fusion requires frame registration to avoid the disturbing artificial effect like a ghost image. The decision level fusion like consensus voting may lead to information loss since it only concerns about the majority. Relatively, feature level is believed to be more effective since it holds sufficient information of all features and avoids the complicated preprocessing of registration owing to its uniform dimension.

Two kinds of traditional fusion schemes are serial [20] and parallel [43] fusion, which are achieved simply by con-

catenating or averaging the features. Although these methods are easy to be implemented, they have a drawback of no consideration of the statistical correlations between pairwise features from different modalities. The average strategy may counteract the strength of one good feature owing to the addition of another one, while the stacking method can cause redundancy and slow down the training process since the high dimension of a fusion feature. In this paper, we blend the features with the canonical correlation analysis (CCA) [30] that tries to maximize the pair-wise correlations across features with different modalities.

Supposing that two matrices $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$ are two features containing n samples from different modalities. The covariance matrix of $\begin{pmatrix} X \\ Y \end{pmatrix}$ can be denoted as:

$$S = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \quad (3)$$

where $S_{xx} \in \mathbb{R}^{p\times p}$, $S_{yy} \in \mathbb{R}^{q\times q}$ denotes the within-set covariance matrices of X and Y, and $S_{xy} \in \mathbb{R}^{p\times q}$ denotes between-set covariance matrix ($S_{xy}^{\mathbf{T}} = S_{yx}$). However, feature vectors from different modalities may not follow a consistent pattern [7], and thus it is hard to obtain the relationships between them through S directly. CCA tries to find a pair of canonical variates with the transformation $X^* = W_x^{\mathbf{T}} X$ and $Y^* = W_y^{\mathbf{T}} Y$, to maximize the pair-wise correlation across two feature sets:

$$Corr(X^*, Y^*) = \frac{cov(X^*, Y^*)}{var(X^*) * var(Y^*)} \quad (4)$$

where $cov(X^*, Y^*) = W_x^{\mathbf{T}} S_{xy} W_y$, $var(X^*) = W_x^{\mathbf{T}} S_{xx} W_x$ and $var(Y^*) = W_y^{\mathbf{T}} S_{yy} W_y$. Lagrange multipliers are used to maximize the covariance between X^* and Y^* with the constraints $var(X^*) = var(Y^*) = 1$. After obtaining W_x and W_y, the blended feature Z, named as Canonical Correlation Discriminant Feature (CCDF), can be performed as:

$$Z = W_x^{\mathbf{T}} X + W_y^{\mathbf{T}} Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^{\mathbf{T}} \begin{pmatrix} X \\ Y \end{pmatrix} \quad (5)$$

4. Experiments

4.1. Dataset

To evaluate the performance, we conduct our experiments on a large-scale RGB-D gesture dataset - the Chalearn LAP IsoGD database. This dataset is built by Wan *et al.*[34], which is derived from the Chalearn Gesture Dataset (CGD) [6]. This dataset is for the task of user independent recognition, namely recognizing gestures without considering the influence of performers. There are 47933 gestures labeling from 1 to 249 that are solely contained in the same amount videos, which can be divided into three subsets: training set (35878 videos), validation set (5784 videos) and testing set (6271 videos). Meanwhile, both RGB and depth data, which is obtained by a Kinect device simultaneously, are available in the dataset.

4.2. Training details

Experimental environment We conduct our experiments on a PC with Intel Core i7-6700 CPU @ 3.40GHz $\times$ 8, 16GB RAM and NVIDIA Geforce GTX TITAN X GPU. The model training and feature extracting are based on the caffe framework [13], and the others are implemented on Matlab R2015b.

Parameter setting The input videos are unified into 32-frame ones and resized to 128×171. Then the video clips are randomly cropped into 112×112. We use SGD with the mini-batches of 2 clips. The starting learning rate is 0.001 and drops at a rate of 0.9 every 5000 iterations. The weight decay and momentum are 0.0005 and 0.9, respectively. The training process is stopped after 120000 iterations.

Batch normalization Batch normalization [11] is widely adopted to accelerate deep network training. It takes a running mean and standard deviation of a mini-batch of input data to achieve normalization and outputs training data with zero mean and unit standard deviation. The BN layer is utilized right after each convolutional layer in our network.

Data augmentation In order to increase the diversity of data, we augment data by following two ways: on the one hand, we use data obtained with different sampling strategies (such as uniform sampling) as a kind of input. On the other hand, as convolutional networks are invariant to rigid transformation, we also mirror the videos for increasing the number of samples.

4.3. Recognition results on different modalities

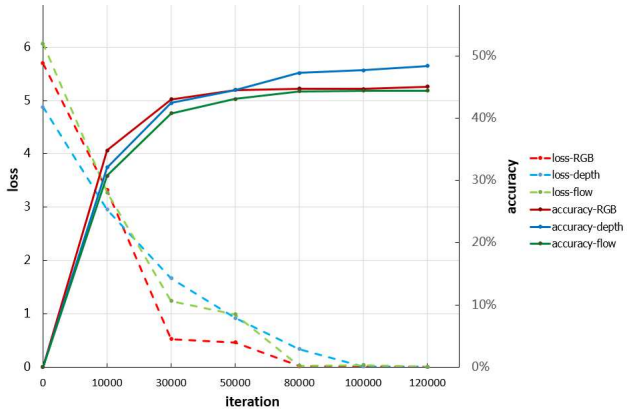


Figure 5. The performance on different modalities along with iteration. The primary y axis (left) indicates the variation of loss while the secondary (right) stands for the accuracy.

In this subsection, we verify the performance of the proposed model on different kinds of data. In this paper, we

utilize three modalities of RGB, depth and optical flow. The recognition results varying along with iterations of these data are shown in Figure 5. As can be seen, in the early stage, the training loss and accuracy change a lot. After about 30000 iterations, both the loss and accuracy tend to be stable. And it shows almost no variation when the iteration reaches 120000.

4.4. Effectiveness of fusion scheme

The effectiveness of our fusion scheme is discussed in this subsection. In addition to the CCA-based fusion method for the multimodal fusion, we also try a fusion of features extracted by different models such as TSN [36].

The comparison across CCDF and individual feature of RGB, depth, and flow, and simply serial and parallel fusions are shown in Figure 6. There is no doubt that the improvement on recognition accuracy of fusion features is significant when compared with the single modality features. That proves the effectiveness of multimodal strategy owing to the sufficient exploitation of comprehensive information. Meanwhile, we can find that the CCA method for fusion also outperforms the other simply fusion scheme to a large extent, which means that a sophisticated analysis of the statistical correlation between pair-wise features from different sets can be of benefit to maximize the correlation between features of the same sample and lead to a better fusion.

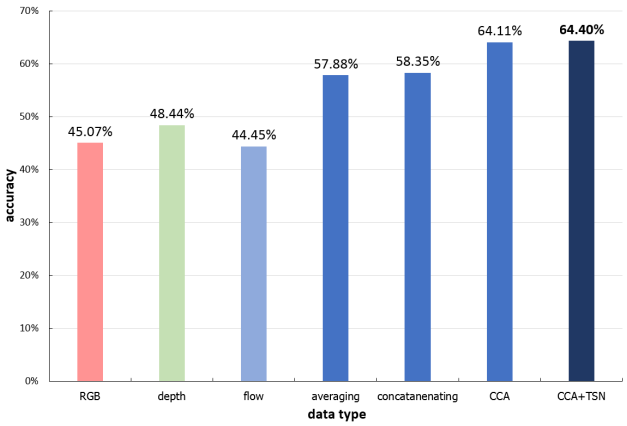


Figure 6. Results of fusion schemes. The accuracy of simple serial (concatenating) and parallel (averaging) fusion schemes together with our CCA based fusion scheme are illustrated and compared with the single modality result. Feature fusion can increase the accuracy by at least 10% and the statistical analysis based scheme achieves better results.

As indicated in [31], the cross-model fusion is a possible way to boost the performance as well. Therefore, we also employ the features extracted by another state-of-the-art method, temporal segment network [36]. As shown in Figure 6, the participant of TSN features is beneficial to recognize gestures.

Method	Accuracy
MFSK+BoVW [34]	18.65%
SFAM (Multi-score Fusion) [37]	36.27%
CNN+depth maps [38]	39.23%
Pyramidal C3D [45]	45.02%
2SCVN+3DDSN [5]	49.17%
32-frame C3D [19]	49.20%
C3D+LSTM [46]	51.02%
proposed method	**64.40 %**

Table 1. Comparisons with state-of-the-art methods in accuracy.

Team	Accuracy (validation)	Accuracy (testing)
baseline [5]	49.17%	67.26 %
XDETVP	58.00%	60.47%
AMRL	60.81%	65.59%
Lostoy	62.02%	65.97%
SYSU_ISEE	59.70%	67.02%
ASU (our)	**64.40 %**	**67.71 %**

Table 2. Final ranking in the Chalearn LAP Large-scale Isolated Gesture Recognition Challenge.

4.5. Comparison with other methods

Comparisons with state-of-the-art methods The performance of our final scheme is compared with several state-of-the-art methods. As only the label of validation set of Chalearn LAP IsoGD database is available, the comparison is conducted on it.

As can be found in Table 1, the neural network based methods witness an improvement of at least 20% on the handcrafted feature based method like [34]. And thanks to the ResC3D network with the combination of a sensible data pre-processing and a statistical analysis based fusion scheme, our method can achieve about 13% accuracy gain on the compound model of C3D and LSTM [46].

Comparisons with entries in the Challenge In Table 2, the result in the ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge (Round 2) [14] is also illustrated. Our method reaches 67.71% on the Chalearn LAP IsoGD database and is the only one that outperforms baseline method [5] on the testing set. This entry wins the 1st place in this challenge.

4.6. Quantitative analysis

The confusion matrix and class-wise recognition rate are shown in Figure 7 and Figure 8, respectively.

The confusion matrix shows that our method can classify the gestures well overall. That can be further demonstrated in the Figure 8 - 30 classes of gestures are fully recognized and over 57% classes are recognized with the accuracy higher than 60%. However, there are still four classes

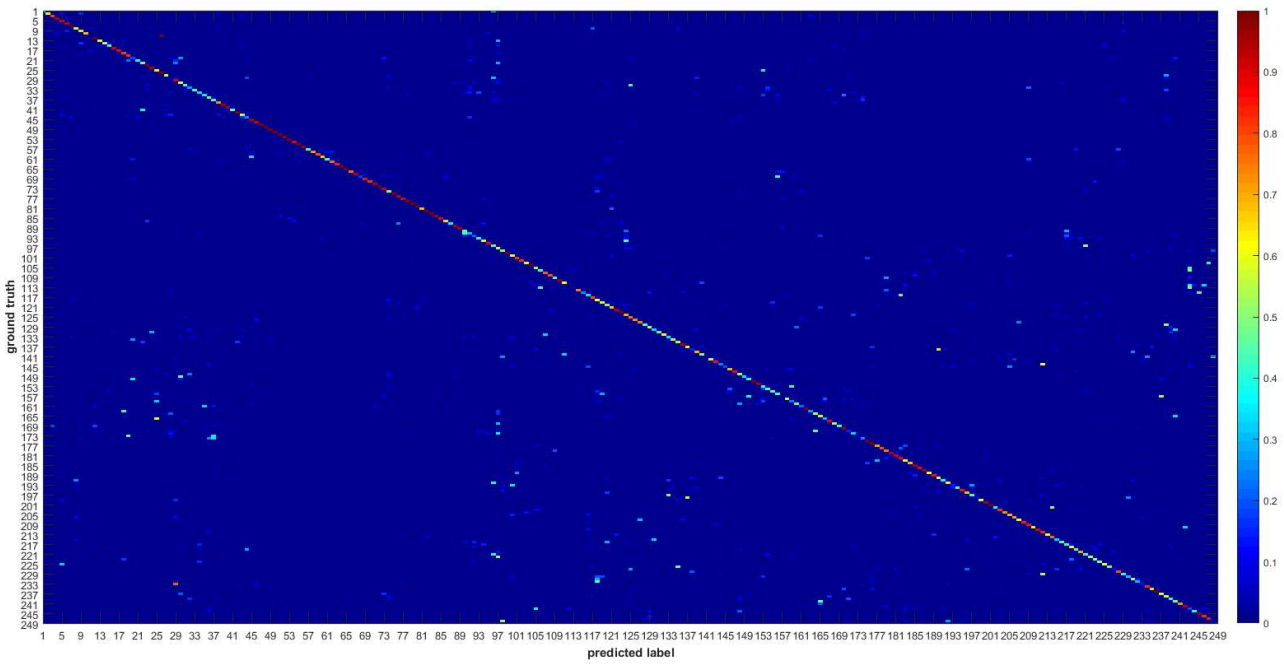


Figure 7. Confusion matrix of our method. The $x-$ and $y-$ axis refer to predicted label and ground truth, respectively. Point at the i_{th} row and j_{th} column represents the sample with label i is classified to j and its color means the proportion of such a condition. Our method yields a nice result on most classes.

are completely wrongly predicted. One reason to account for is some of the gestures are too difficult to distinguish. Take the wrongly classified gesture 11 as an example, they are almost all the same as gesture 26 from the motion path to the final posture as shown in Figure 9. There is also an interesting phenomenon that only gestures with label of 11 are confused with 26, whereas the gesture 26 are all recognized correctly. That may because the SVM classifier we adopt is composed of 249 binary classifiers. That will update for each label and consequently, the gesture 11 are all given the label of 26.

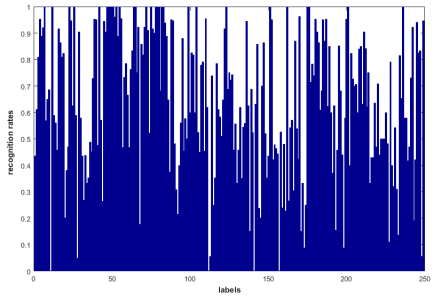
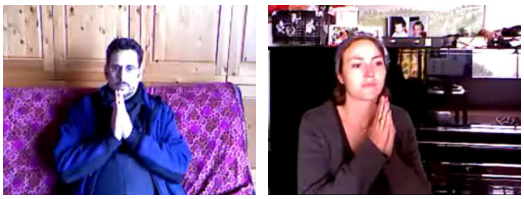


Figure 8. Per-class recognition rate. Over 57% classes are recognized with the accuracy higher than 60%. Nevertheless, there are still four classes completely confused.



(a) (b)

Figure 9. The wrongly classified gestures. (a) gesture 11. (b) gesture 26.

5. Conclusion

In this paper, we present a multimodal gesture recognition method. We first eliminate the gesture-irrelevant factors, such as illumination and noise in the RGB and depth data. After that, we propose a key frame attention mechanism and based on that we select the most representative frames in a video. Following we extract features by the ResC3D network and then blend them with canonical correlation analysis. The final recognition result is derived with a linear SVM classifier. We achieve 67.71% accuracy on the testing set of Chalearn LAP IsoGD dataset and win the 1st place in the isolated gesture recognition challenge.

However, there is still some room for improving the recognition performance. One way is to find more sophisticated feature to distinguish gestures with subtle differ-

ences. Meanwhile, the influence of gesture-irrelevant factors caused by the performers, like the velocity of movement or skin color, needs to be eliminated to better concentrate on the gesture itself.

Acknowledgement

The work was jointly supported by the National Natural Science Foundations of China under grant No. 61472302, U1404620 and 61672409; The Open Projects Program of National Laboratory of Pattern Recognition(201600031); The Fundamental Research Funds for the Central Universities under grant No.JB150317; Natural Science Foundation of Shaanxi Province, under grant No. 2010JM8027. The Aviation Science Foundation under Grant No. 2015ZC31005.

References

[1] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, pages 25–36. Springer, 2004.

[3] A. Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *IEEE International Conference on Computer Vision Workshops*, pages 82–89. IEEE, 2001.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2625–2634, 2015.

[5] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Li. A unified framework for multi-modal isolated gesture recognition. In *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),(Accept)*, 2017.

[6] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante. The chalearn gesture dataset (cgd 2011). *Machine Vision and Applications*, 25(8):1929–1951, 2014.

[7] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Transactions on Information Forensics and Security*, 11(9):1984–1996, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 770–778, 2016.

[9] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 410–415. IEEE, 2000.

[10] Y. Huang, Y. Gao, H. Wang, D. Hao, J. Zhao, and Z. Zhao. Enhancement of ultrasonic image based on the multi-scale retinex theory. In *Recent Advances in Computer Science and Information Engineering*, pages 115–120. Springer, 2012.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[12] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[14] W. Jun, S. Escalera, A. Gholamreza, H. J. Escalante, X. Baró, I. Guyon, M. Madadi, A. Juri, G. Jelena, L. Chi, and X. Yiliang. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *IEEE International Conference on Computer Vision Workshops*, 2017.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[16] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 1–10. British Machine Vision Association, 2008.

[17] J. Konecnỳ and M. Hagara. One-shot-learning gesture recognition using hog-hof. *Journal of Machine Learning Research*, 15:2513–2532, 2014.

[18] E. H. Land and J. J. McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[19] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *IEEE International Conference on Pattern Recognition Workshops*. IEEE, 2016.

[20] C. Liu and H. Wechsler. A shape-and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions On Image Processing*, 10(4):598–608, 2001.

[21] M. R. Malgireddy, I. Inwogu, and V. Govindaraju. A temporal bayesian model for classifying, detecting and localizing activities in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 43–48. IEEE, 2012.

[22] M. R. Malgireddy, I. Nwogu, and V. Govindaraju. Language-motivated approaches to action recognition. *Journal of Machine Learning Research*, 14(1):2189–2212, 2013.

[23] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215. IEEE, 2016.

[24] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *IEEE International Conference*

on Signal and Image Processing Applications Workshops, pages 342–347. IEEE, 2011.

[25] K. Oka, Y. Sato, and H. Koike. Real-time fingertip tracking and gesture recognition. *IEEE Computer graphics and Applications*, 22(6):64–71, 2002.

[26] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.

[27] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *IEEE Workshops on Applications of Computer Vision*, pages 103–110. IEEE, 2013.

[28] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[30] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia. A new method of feature fusion and its application in image recognition. *Pattern Recognition*, 38(12):2437–2448, 2005.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, pages 4489–4497. IEEE, 2015.

[32] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[33] J. Wan, G. Guo, and S. Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1626–1639, 2015.

[34] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64. IEEE, 2016.

[35] J. Wan, Q. Ruan, W. Li, G. An, and R. Zhao. 3d smosift: three-dimensional sparse motion scale invariant feature transform for activity recognition from rgb-d videos. *Journal of Electronic Imaging*, 23(2):3017–3017, 2014.

[36] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[37] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *IEEE IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[38] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. In *IEEE International Conference on Pattern Recognition Workshops*, 2016.

[39] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 2, pages 1521–1527. IEEE, 2006.

[40] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.

[41] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385. IEEE, 1992.

[42] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *IEEE International Conference on Computer Vision*, pages 4633–4641, 2015.

[43] J. Yang and J.-y. Yang. Generalized k–l transform based combined feature extraction. *Pattern Recognition*, 35(1):295–297, 2002.

[44] M. Yeasin and S. Chaudhuri. Visual understanding of dynamic hand gestures. *Pattern Recognition*, 33(11):1805–1817, 2000.

[45] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *IEEE International Conference on Pattern Recognition Workshops*, 2016.

[46] G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3d convolution and convolutional lstm. *IEEE Access*, 2017.