EMOTION RECOGNITION BASED ON MULTI-VIEW BODY GESTURES

Zhijuan Shen[1,2,3], Jun Cheng[2,3,], Xiping Hu[2,3], Qian Dong[2]*

[1]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3]The Chinese University of Hong Kong, Hong Kong, China

ABSTRACT

Body gesture, a crucial component of "body language", remains less explored to recognize emotion while face expression-based and speech-based approaches are widely investigated. In this paper, we introduce a multi-view body gesture dataset (named as MBGD) for emotion recognition, which consists of 43200 RGB videos of simplified body gestures of six basic emotions and their neutral control groups. This dataset is large scale and multi-view to use data-driven methods like deep learning and various evaluations. Furthermore, a novel approach is proposed to fuse skeleton and RGB multi-modality only using single-modality RGB video data. Experimental results show our approach achieves substantial improvements both in individual categories and overall and has stronger generalization capability as well.

Index Terms— emotion recognition, body gesture, multi-view, feature fusion

1. INTRODUCTION

Human emotion recognition is an active research topic involving many applications such as users' emotions detection, actual or virtual animated conversational agents designing, mental fitness monitoring, etc. In the past decade, while works of emotion recognition based on facial expressions or speech abounded, recognizing affect from body gestures remains a less explored topic. However, body gesture is one of the most crucial components of what is commonly known as "body language", which conveys nonverbal cues even when the face and voice unavailable.

Referring to the survey [1], the quantity of labelled emotional body gesture dataset is scarce. All available datasets are summarized in Table 1. On account of datasets limitation, emotion recognition from body gesture largely based on naive geometrical representations, such as displacements [8], orientation of hands [6] and motion cues like velocity and acceleration [9]. Conventional researches mainly focus on hand-craft features, and the machine learning algorithms, such as SVM, Naive Bayes, RF [10], and DTW [11], are the most popular classification methods which have been used in these approaches.

Table 1. Available Body Gesture Datasets for Emotion Recognition

Name	Body parts	Emotions	Subjects	Samples
FABO [2]	Face and body	10	23	206
GEMEP [3]	Face and body	18	10	>7000
HUMAINE [4]	Face and body	8	10	240
LIRIS-ACCEDE [5]	Face and upper body	6	64	NA
THEATER [6]	Body	8	NA	NA
EMILYA [7]	Body	8	11	NA

To overcome the existing limitations and apply the advanced data-driven learning methods to this aspect, we collect a multi-view body gesture dataset for emotion recognition, named as MBGD. It consists of simplified body gestures for 6 basic types of emotions and their neutral control groups, captured from 80 human subjects (40 females and 40 males) using Hikvision network cameras from 15 views and repeat 3 times, generating totally 43200 samples.

In addition, inspired by [12], we propose a novel approach based on deep neural networks and transfer learning for this task. Our approach fuse RGB and skeleton features only using RGB videos without motion-capture devices. Experiments show that the proposed framework can substantially improve accuracies both in individual categories and overall on cross-subject and cross-view evaluations. It is a heuristic attempt, because RGB cameras are more widely available and lower cost than motion-capture devices. The code of our approach is made publicly available[1].

The remainder of this paper is organized as follows. Section 2 introduces our dataset, MBGD. Section 3 presents the details of our approach. Section 4 shows the experimental evaluations, and conclusion and future work are given in section 5.

[1] https:// github.com/DQ0408/MBGE-recognition.

2. MULTI-VIEW BODY GESTURE DATASET

2.1. Dataset description

Body gesture categories. Body gestures, which differ from facial expressions and speech, do not have obvious emotional traits. For the same emotion, different people express different body gestures. Even if the same individual performs diverse body gestures vary from situations. This paper as beginning of our research, we simplify the gesture to reduce the difficulty of recognition. In the dataset, we choose 6 basic types of emotions (happiness, sadness, anger, surprise, fear and disgust) [13] as well as their neutral control groups as research objects. Samples of body gestures in our dataset are illustrated in Figure 1.

Subjects, sensors and data modality. To collect this dataset, we invited 80 distinct subjects (40 females and 40 males). The ages of the subjects are between 17 and 31. We utilized Hikvision network cameras to record RGB videos in the provided resolution of 1280x720 pixels and sampling rate of 25 frames per second. Our dataset, which has 43200 RGB video samples collected from 80 human and 15 different views, is a strong complement to now available body gesture datasets. This large amount of variation in subjects and views makes it competent to data-hungry methods such as deep learning and to have more accurate cross-subject and cross-view evaluations.

2.2. Benchmark evaluations

We utilized 15 cameras to record simultaneously. Shown in Figure 2, each camera is assigned a consistent ID number. For each setup, the 15 cameras were located at 3 different height (top view, head view, and bottom view) and from 12 different horizontal angles, which look like a clock dial.

To conduct comparable evaluations for existing and future works on this benchmark, we define precise criteria for classification evaluation, i.e., cross-subject evaluation and cross-view evaluation, as described in this section. For each evaluation, we report the classification accuracy in percentage.

Cross-subject evaluation. In cross-subject evaluation, we split the 80 subjects into training (60%) and testing (40%) groups. For this evaluation, the training and testing sets have 25920 and 17280 samples, respectively.

Cross-view evaluation. We propose two cross-view evaluations, one by rotation angle and the other by angle of pitch. Shown in Figure 3, for cross-view evaluation 1, we pick all the samples of camera 5,8,9,14,15 for testing and others for training. For this evaluation, the training and testing sets have 28800 and 14400 samples, respectively. For cross-view evaluation 2, we pick all the samples of head-view and bottom-view cameras for training and top-view for testing. For this evaluation, the training and testing sets have 25920 and 17280 samples, respectively.

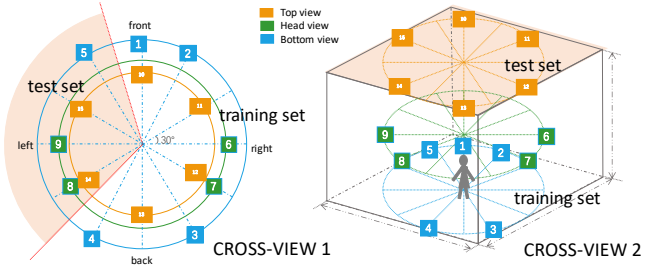


Fig. 2. The multiple views of MBGD

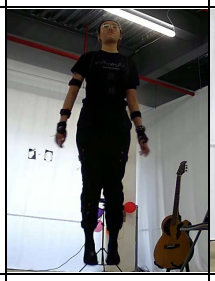
Body gestures	Jump	Squat	Throw	Stand	Recede	Turn and walk away
Neutral						
Emotional	Happiness	Sadness	Anger	Surprise	Fear	Disgust
						

Fig. 1. The body gestures in MBGD

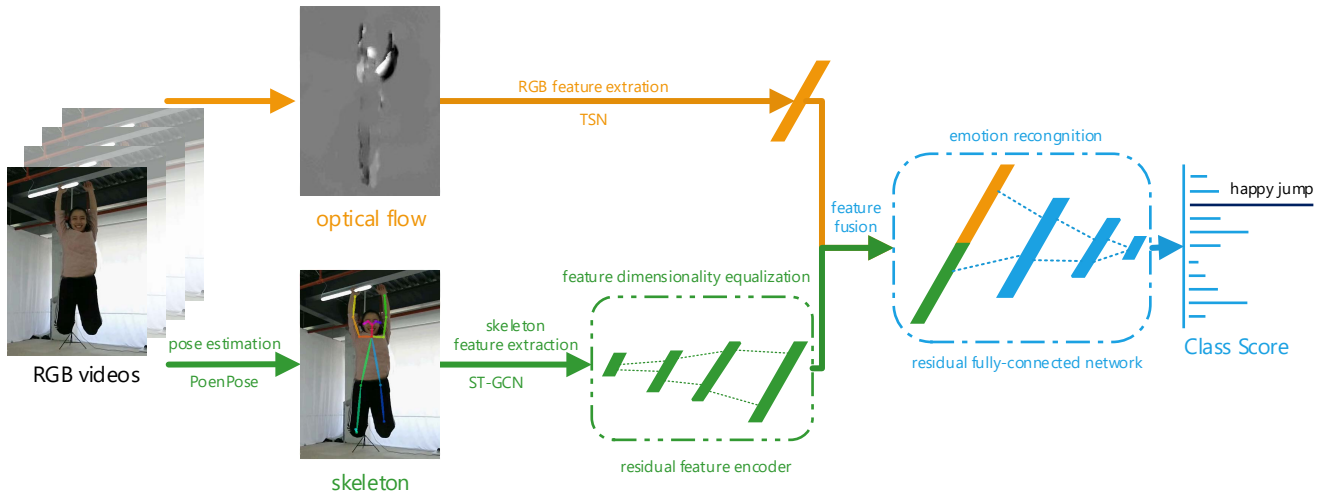


Fig. 3. Pipeline of our approach

3. THE PROPOSED METHOD

3.1. Pipeline Overview

According to [12], acceleration for hand is the base classifier of ensemble tree classifier, the model which performance best in their study. We realize that different from actions, emotional body gestures are more closely related to velocity and acceleration of key joints. For instance, when a person is angry, the speed of throwing things by hand becomes faster.

In our approach, we use the publicly available OpenPose [14] toolbox to estimate the location of the joints first. In order to get RGB feature and skeleton feature, TSN [15] and ST-GCN [16], two advanced action recognition models, are used separately. Outputs of them are vectors unequal in length. To avoid contribution inequality, we design a residual feature encoder to make them equal. The two feature vectors are concatenated to a new fusion feature and be classified by a full-connected residual network into the corresponding emotional body gesture category.

3.2. Feature extraction

An amount of samples and variation in and views make it possible to use data-hungry deep learning models to extract feature rather than hand crafting.

RGB feature. TSN (temporal segment network) [16], a state-of-the-art model for action recognition, is used to extract the RGB feature. The output is a 1024-dimensional RGB feature vector.

Skeleton feature. Skeleton-based data can be obtained from motion-capture devices or pose estimation algorithms from videos. We only have raw RGB videos, so the public available OpenPose [15] toolbox is used to estimate the location of joints. Then ST-GCN (Spatial-Temporal Graph Convolutional Networks) [17], a model based on skeleton using graph ConvNets is used to extract the skeleton feature. The output is a 400-dimensional RGB feature vector.

3.3. Feature dimensionality equalization and fusion

To avoid contribution inequality, we design a residual feature encoder to ascend skeleton feature vector to 1024-dimensional. The encoder is a residual fully-connected network and has the typical structure of ResNet in [17]. There are 3 same-structure basic blocks in the encoder. Each has 3 full-connection layers followed by a PReLU activation layer and drop out layer to avoid over-fitting. The outputs of the first full-connection layer and third have the same dimensions, and there is a shortcut connection to which turn the network into its counterpart residual version. As for weight optimization, we use Cross-entropy Method.

Then we concatenate the two 1024-dimensional features into a 2048-dimensional feature vector.

3.4. Emotion recognition

We use a residual fully-connected network, which has the same structure as the residual feature encoder mentioned above, to categorize the emotional body gestures.

4. EXPERIMENTS

4.1. Experimental setup

As shown in Table 1 in instruction, the quantity of labeled emotional body gesture dataset is scarce, and only 6 available. And the number of video samples in each dataset is limited to apply deep learning, we only performed experiments on our dataset.

In order to train TSN to achieve optimal performance, we studied a series of practical matters. After comparing all the results, we chose optical flow as input, BN-Inception as

training setting, and AVG as consensus function. As for ST-GCN, we estimated 18 joints as input by OpenPose. For multi-person cases by estimation error, we select the person with the highest average joint confidence given by OpenPose.

All experiments were conducted on PyTorch deep learning framework with 4 GeForce GTX 1080Ti GPUs.

4.2. Experimental Evaluations

The experimental results are evaluated from both dataset aspect and approach aspect.

Dataset aspect. According to the confusion matrix in Figure 4, the recognition accuracies of all categories are high basically. It means that our dataset can be recognized effectively using appropriate approach, even though the difference between the control group is small. The accuracy of happiness, anger and surprise is lower, because they are passion emotions (emotion can be divided into mood, passion and stirring emotions), which are speedy and varies from man to man. For example, when feeling surprised, some people will raise their hands and cover their mouths, while others will just scream without body gesture changing.

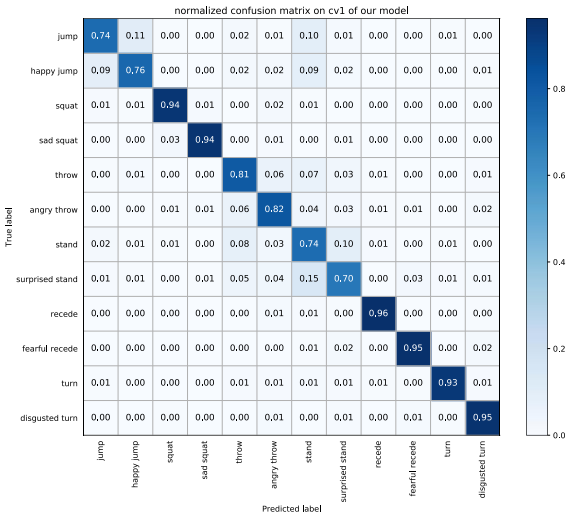


Fig. 4. The confusion matrix of our approach on cv1

Approach aspect. Comparing the first two rows of Table 2, ST-GCN performs better than TSN. Because different from actions, emotional body gestures are more closely related to velocity and acceleration of key joints and ST-GCN is based on the skeleton. In addition, skeletal representation is view-invariant, which makes ST-GCN stronger to generalize between the views, but it's prone to errors of pose estimation.

As shown in Table 2 and Figure 5, the recognition accuracies of our approach are substantially improved over TSN and ST-GCN both in individual categories and overall. It means that the skeleton from pose estimation rather than motion-capture devices, can also complement RGB videos to perform better in recognition. In particular, the third row of Table 2 shows that our approach performs well both in cross-subject and cross-view evaluations. Our approach not only leads to greater recognition power but also stronger generalization capability.

Table 2. The overall accuracies of TSN, ST-GCN and our approach

	Cross-subject	Cross-view1	Cross-view2
TSN	70%	72.09%	61.55%
ST-GCN	75.32%	72%	72.61%
Our approach	**85.47%**	**85%**	**82.31%**

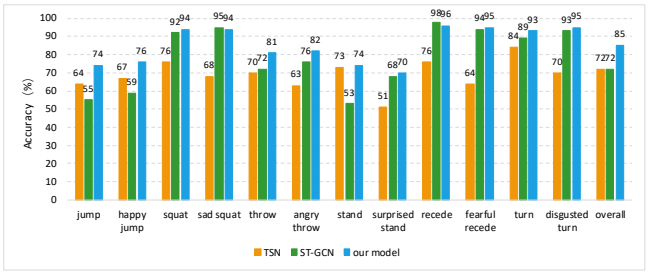


Fig. 5. The category and overall accuracies of our approach on cross-view1

5. CONCLUSION AND FUTURE WORK

A multi-view body gesture dataset (name as MBGD) for emotion recognition is introduced in this paper. Our dataset, which has 43200 RGB video samples collected from 80 human and 15 different views, is a strong complement to now available body gesture datasets. This makes it possible to apply data-hungry methods such as deep learning and to have more accurate cross-subject and cross-view evaluations.

In addition, we propose a novel approach to fuse skeleton and RGB multi-modality only using single-modality RGB video data. Compared with motion-capture devices, RGB cameras are more widely available and lower cost. We hope our approach to boost research in multi-modality fusion from fewer modalities by software programming, independent of hardware limitations.

The provided experimental results show the availability and effectiveness of our approach, which can substantially improve accuracies of recognition both in individual categories and overall. Evaluations of cross-subject and cross-view show that our approach has strong generalization capability.

In future work, we will do more research both in standard emotional body gesture dataset and improve our model construction to achieve better performance with less computing resources and time.

6. REFERENCES

[1] C.A. Corneanu, F. Noroozi, D. Kami´nska, T. Sapiński, S. Escalera and G. Anbarjafari, "Survey on emotional body gesture recognition," *arXiv1801.07481*, 2018.

[2] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior", *18th International Conference on Pattern Recognition*, pp. 1148-1153, 2006.

[3] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-6,2008.

[4] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," *International Conference on Affective Computing and Intelligent Interaction*, pp. 71-82, 2007.

[5] M. Gavrilescu, "Recognizing emotions from videos by studying facial expressions, body postures and hand gestures," *Telecommunications Forum*, pp. 720-723, 2015.

[6] M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?" *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1-8, 2009.

[7] N. Fourati and C. Pelachaud, "Emilya: Emotional body expression in daily actions database," *LREC*, pp. 3486-3493, 2014.

[8] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 3437-3443, 2005.

[9] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. J. I. T. o. A. C. Scherer, "Toward a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, pp. 106-118, 2011.

[10] S. Li, L. Cui, C. Zhu, B. Li, N. Zhao, and T. J. P. Zhu, "Emotion recognition using Kinect motion capture data of human gaits," *PeerJ*, vol. 4, pp.e. 2336, 2016.

[11] J. Arunnehru, and M. K. Geetha, "Automatic human emotion recognition in surveillance video," *Intelligent Techniques in Signal Processing for Multimedia Security*. Springer, Cham, pp. 321-342,2017.

[12] S. Saha, S. Datta, A. Konar, and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," *International Conference on Communication and Signal Processing*, pp. 56-60, 2014.

[13] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique", *Psychological Bulletin*, pp. 268-287,1994.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299, 2017.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, et al., "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] S. Yan, Y. Xiong, and D. J. a. p. a. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *arXiv:1801.07455*, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun., "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.