Facial Expression Recognition Based on Fused Spatio-temporal Features

Jingjie Yan
The School of Information Science and Engineering,
Southeast University
Nanjing, China
E-mail: yanjingjie1212@163.com

Minghan Xin
Research Center for
Learning Science, Southeast University
Nanjing, China
E-mail: xinminghai@163.com

***Abstract*—Although spatio-temporal features (ST) have recently been developed and shown to be available for facial expression recognition and behavior recognition in videos, it utilizes the method of directly flattening the cuboid into a vector as a feature vector for recognition which causes the obtained vector is likely to potentially sensitive to small cuboid perturbations or noises. To overcome the drawback of spatio-temporal features, we propose a novel method called fused spatio-temporal features (FST) method utilizing the separable linear filters to detect interesting points and fusing two cuboids representation methods including local histogrammed gradient descriptor and flattening the cuboid into a vector for cuboids descriptor. The proposed FST method may robustness to small cuboid perturbations or noises and also preserve both spatial and temporal positional information. The experimental results on two video-based facial expression databases demonstrate the effectiveness of the proposed method.**

Keywords- facial expression recognition; spatio-temporal (ST); fused spatio-temporal features (FST)

I. INTRODUCTION

The recognition of human's facial expression has become one of the most interesting and popular research topic, and impacts many important applications such as human computer interaction (HCI). Generally, most of the facial expression recognition methods attempt to recognize a given facial image or video into six basic emotion categories defined by Ekman and Friesen [1], i.e. angry, disgust, fear, happy, sad, and surprise. During the last decades, facial expression recognition had been widely studied and many different approaches had been presented in the literatures. For a literature survey, see [2] [3]. Although much progress has been acquired, recognizing facial expression in video sequences remains difficult because of the complexity and variability of facial expressions.

In handling the facial expression recognition in video sequences, the first step is to derive available features from original video sequences. One major problem we encounter with is the high dimension of facial image sequences, and it is very possible that the features of facial image sequences carry useless discriminative information and redundant information or noise. So the feature extraction and selection is the key to facial expression recognition in video sequences and it would be helpful to reduce the computational complexity while improve the recognition rate.

Obtaining a set of features which comprise the description of motion of facial expression in the video sequences is a vital step for efficient facial expression recognition. Spatio-temporal (ST) have recently been developed and shown to be available for facial expression recognition, human action categorization and behavior recognition in videos. Efros et al. [4] proposed a spatio-temporal descriptor based on optical flow measurements in a spatio-temporal volume, which was applied to recognize actions on three sport datasets (ballet, tennis and football). Laptev and Lindeberg [5] presented a spatio-temporal of Harris interest point detector by extending the spatial interest points into the spatio-temporal domain, and then applied to recognize human action. Dollar et al. [6] proposed a new spatio-temporal feature detector for the behavior recognition based on separable linear filters. Ryoo et al. [7] extended the use of spatio-temporal features to the recognition of multi-person activities like handshake, push, kick and punch by the analysis of spatial-temporal relationships. Rapantzikos et al. [8] extended the cuboid features to color and motion information as well.

Dollar's spatio-temporal features have been proven available for facial expression recognition and behavior recognition in videos. The method of spatio-temporal features makes little assumptions with facial image sequences compared to some other methods [9] [10] for facial expression recognition in video, such as background, occlusion and so on. But spatio-temporal features has one major disadvantage that the obtained feature vector for representation of cuboids is likely to potentially sensitive to small cuboid perturbations and noises due to utilizing the method of directly flattening the cuboid into a vector as a feature vector.

In this paper, we investigate the emotion recognition of facial expression in video sequences. To overcome the drawback of the spatio-temporal features, we propose a fused spatio-temporal features (FST) method for representing facial expression in video sequences. The fused spatio-temporal features utilizes the separable linear filters to detect interesting points and fuses two cuboids representation methods involve local histogrammed gradient descriptor and flattening the cuboid into a vector for cuboids descriptor. The proposed FST method is robustness to small cuboid perturbations or noises and also preserves both spatial and temporal positional information compared to the ST method.

Published by Atlantis Press, Paris, France.

II. SPATIO-TEMPORAL FEATURES

In this section, we study the spatio-temporal features method to describe facial expression in videos. The method of spatio-temporal features is suit for ascertaining spatially region whose spatially different characteristics experiencing a complex or strong motion.

Spatio-temporal features is based on separable linear filters and was proposed by Dollar et al. in 2005 [6]. In the following, we provide a brief review of the spatio-temporal features method. For a stack of images in the video denoted by $\mathrm{I}(x, y, t)$ the response function has the form:

$$R = \left(I * g * h_{ev}\right)^2 + (I * g * h_{od})^2$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel applied on the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied on temporally. These are given by $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$ where $\omega = 4/\tau$ and the two parameters σ and τ are the spatial and temporal scale of the cuboid detector.

Each space-time interest point is picked up as the local maxima of above response function. It was noted in [6] that any region whose spatially different characteristics experiencing a complex or strong motion can induce a strong response. At each interest point, a cuboid is picked up which comprises the most near neighbour spatio-temporal windowed pixel values. The size of the each cuboid is set as about six times the scale along three dimensions such that it can cover most of the data which contributed to the response function at each interest point. Figure 1 shows examples of cuboids picked up from the FABO database by the spatio-temporal features methods.

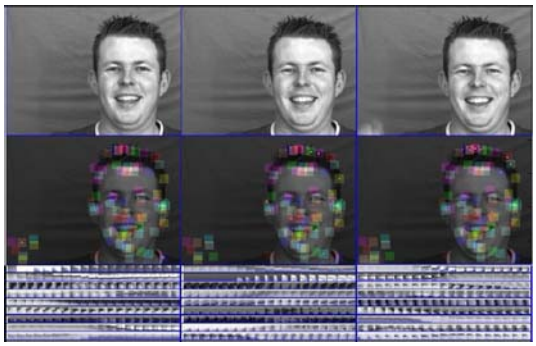

Figure 1. Examples of spatio-temporal features method picked up from the FABO database: the first row is the original facial video sequences; the second row is the visualization of the extracted cuboids; the third row is cuboids flattened with time dimension.

Three cuboids conversion methods are applied to the cuboids including normalized pixel values, brightness gradient and windowed optical flow [6]. As suggested, the flattened gradient obtains the best performance among three cuboids conversion methods. This cuboids conversion method is then projected to a lower dimensional subspace space through the principal component analysis (PCA). At last we can obtain a histogram vectors after clustering all cuboids.

III. FUSED SPATIO-TEMPORAL FEATURES

In the flowing section, we present the fused spatio-temporal features method to represent facial expression in videos. The fused spatio-temporal features method also utilizes the separable linear filters to detect interesting points like spatio-temporal features method. At each interest point, a cuboid is extracted which covers the most near neighbour spatio-temporal windowed pixel values. After obtaining many cuboids, we should use some methods to create a feature vector of the cuboids. The spatio-temporal features utilizes the method of directly flattening the cuboid into a vector which causes the obtained vector is possiblely to sensitive to cuboid perturbations or noises. So we propose the fused spatial-temporal features (FST) method which fuses two cuboids representation methods including local histogrammed gradient descriptor and flattening the cuboid into a vector for cuboids descriptor to overcome the drawback of spatio-temporal features method.

The local histogrammed gradient descriptor is motivate by [11]. For a cuboid extracted in the original video named by $\mathrm{M}(x, y, t)$, the linear scale-space representation L is expressed as the convolution of M with a spatio-temporal Gaussian kernel whose spatial and temporal parameters are σ and τ , i.e.,

$$L(x, y, t; \sigma^2, \tau^2) = g\left(x, y, t; \sigma^2, \tau^2\right) * M(x, y, t),$$

where the spatio-temporal separable Gaussian kernel is defined as

$$g\left(x, y, t; \sigma^2, \tau^2\right) = 1/\sqrt{(2\pi)^3\sigma^4\tau^2} \times \exp(-(x^2 + y^2)/2\sigma^2 - t^2/2\tau^2).$$

then we calculate its gradients on three dimensions

$$L_x\left(x, y, t; \sigma^2, \tau^2\right) = \partial_x\left(x, y, t; \sigma^2, \tau^2\right)$$
$$L_y\left(x, y, t; \sigma^2, \tau^2\right) = \partial_y\left(x, y, t; \sigma^2, \tau^2\right)$$
$$L_t\left(x, y, t; \sigma^2, \tau^2\right) = \partial_t\left(x, y, t; \sigma^2, \tau^2\right)$$

To create a descriptor for cuboids which is robust to small perturbations of the cuboid, we use the local histograms which regarded as part of SIFT descriptor [11]. Every cuboid is classified into equally sized regions which are overlapped aim to cover the cuboid completely. Then each region creates six dependent histograms of the region. At last we can obtain a set of position dependent histograms of the values in every cuboid.

The local histogrammed gradient descriptor for the cuboid is robust to perturbations and noise but discards all spatial and temporal positional information, and the method of flattening the cuboid into a vector is sensitive to small perturbations and noise but retains positional information. To overcome the each drawback of the two methods, we concatenate the respective obtained descriptor to be a vector which is robustness to small perturbations while retains some positional information. Such expresssion can effectively represent available information of facial expression emotion recognition in videos. To reduce the dimensionality, the obtained combination vector is projected to a lower

dimensional subspace space through PCA. We derive a library of cuboid prototypes by clustering many cuboids extracted from the facial video using the k-means algorithm. Then each cuboid can be obtained a category by mapping it to the similar prototype vector. At last, we also use a histogram of the cuboid category as features which are the input to the next recognition procedure.

IV. EXPERIMENTS

In this section, we use two video-based facial expression databases to evaluate the performance of the proposed method. The detail of two databases is introduced in the following section. We use "leave one subject out" cross-validation [12] strategy in the experiments of two databases. We utilize the Support Vector Machine (SVM) classifier and the nearest neighbor classifier as classifier for facial expression recognition.

A. Experiments on Dollar's Facial Expression Database

Dollar's facial expression database \cite{Dollar} is created by Dollar, and it can be download at http://vision.ucsd.edu. It is a small video-based facial expression database which has only two individuals and six different basic emotions, i.e. angry, disgust, fear, joy, sadness, and surprise. Figure 2 shows some sample images from the Dollar's database.

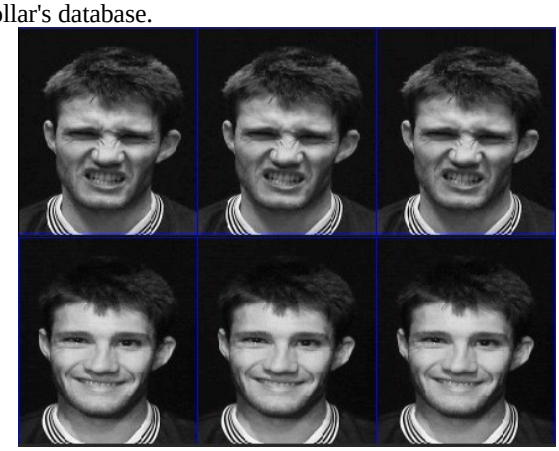

Figure 2. Examples of facial expression from the Dollar's database.

We compare the recognition rate of the fused spatio-temporal features (FST) with the spatio-temporal features (ST) using the SVM classifier and the nearest neighbor classifier for facial expression recognition respectively. Figure 3 and Figure 4 show the confusion matrices of two feature extraction methods with the SVM classifier respectively. The average recognition rate of two feature extraction methods with the SVM classifier and the nearest neighbor classifier is displayed in Table 1.

B. Experiments on FABO Database

Gunes and Piccardi create the FABO database which consists of facial expression and body gesture recorded simultaneously [13]. Figure 5 shows sample images from the FABO database.

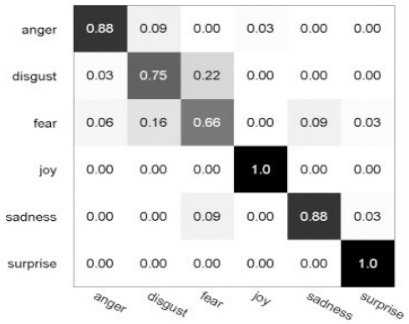

Figure 3. Confusion matrices of facial expression recognition with the spatio-temporal features.

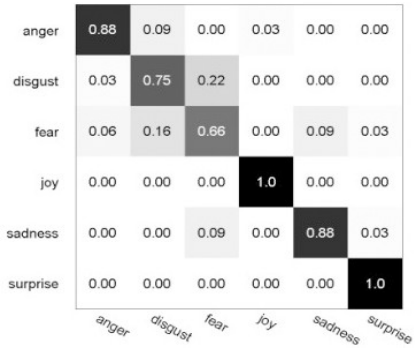

Figure 4. Confusion matrices of facial expression recognition with the fused spatio-temporal features.

TABLE I. THE AVERAGE RECOGNITION RATE OF TWO FEATURE EXTRACTION METHODS WITH SVM CLASSIFIER AND KNN CLASSIFIER

Method	KNN	SVM
ST	80.21	83.85
FST	83.33	85.94

Figure 5. Examples of facial expression from the FABO database.

The FABO database contains over 1900 videos from 23 subjects aging from 18 two 50. In our experiments, a sub-database of facial expression video is used. Specifically, we select 89 videos of four emotions (Boredom, Disgust, Happiness and Uncertainty) from 10 subjects. In order to reduce computational complexity, the data resolution of original video (1024*768 pixels) are automatically down-sampled to 256*192 proportionally.

We also compare the recognition rate of the fused spatio-temporal features (FST) with the spatio-temporal features

(ST) using the SVM classifier and the nearest neighbor classifier for facial expression recognition respectively. Figure 6 and Figure 7 show the confusion matrices of two feature extraction methods with the SVM classifier respectively. The average recognition rate of two feature extraction methods with the SVM classifier and the nearest neighbor classifier is given in Table 2.

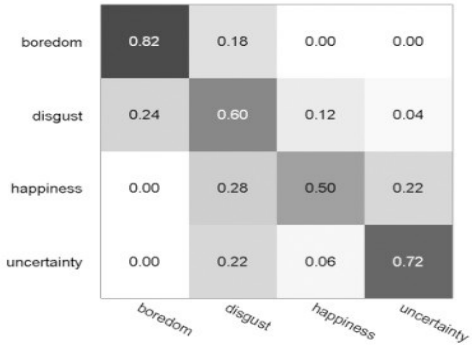

Figure 6. Confusion matrices of facial expression recognition with the spatio-temporal features.

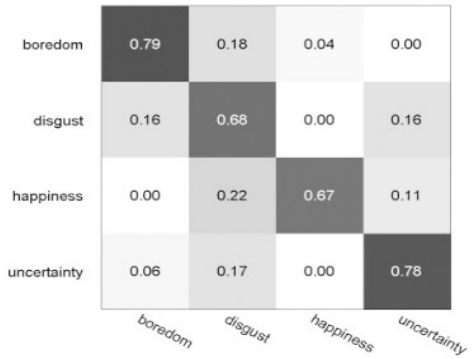

Figure 7. Confusion matrices of facial expression recognition with the fused spatio-temporal features.

TABLE II. THE AVERAGE RECOGNITION RATE OF TWO FEATURE EXTRACTION METHODS WITH SVM CLASSIFIER AND KNN CLASSIFIER

Method	KNN	SVM
ST	58.43	67.42
FST	60.67	73.03

From the above results, we can see that the recognition of fused spatio-temporal features (FST) is better than spatio-temporal features (ST) on both Dollar's and FABO facial expression databases. Especially, for the FABO database, the FST method is much higher than the ST method using the SVM classifier. This is possibly because the FST method is not only robustness to small cuboid perturbations or noises but also preserve both spatial and temporal positional information, so it keep more effective discriminant information and improve the performance of facial expression recognition.

V. CONCLUSIONS

In this paper, we introduce a novel video-based emotion recognition approach from facial expression. We present a fused spatio-temporal features method which utilizes the separable linear filters to detect interesting points and fuses two cuboids representation methods involve local histogrammed gradient descriptor and flattening the cuboid into a vector for cuboids descriptor. The proposed FST method is robustness to small cuboid perturbations or noises and also preserve both spatial and temporal positional information compared to the ST method. The experimental results on Dollar's and FABO facial expressions databases demonstrate the improvement of the proposed method.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under Grant No. 61231002, No. 61273266 and No. 51075068, Doctoral Fund of Ministry of Education of China under Grant No. 20110092130004.

REFERENCES

[1] P. Ekman and W. V. Friesen, “Pictures of facial affect,” in Human Interaction Laboratory, San Francisco, CA: Univ. California Medical Center, 1976.

[2] Y.L. Tian, T. Kanade, J.F. Cohn, “Facial Expression Analysis, ” In: S.Z. Li, A.K. Jain (eds.), Handbook of Facial Recognition, Springer, New York, USA, 2005, pp.247-276.

[3] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, ” IEEE TPAMI, 2009, Vol.31, No.1, pp.39-58.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, “Recognizing action at a distance, ” In IEEE International Conference on Computer Vision (ICCV) , 2003, pp.726-733.

[5] I. Laptev and T. Lindeberg, “Space-time interest points,” In IEEE International Conference on Computer Vision, 2003, pp.432-439.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatiotemporal features,” In VS-PETS, 2005, pp.65-72.

[7] M. S. Ryoo, and J. K. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” In: International Conference on Computer Vision, 2009, pp.1593-1600.

[8] K. Rapantzikos, Y. Avrithis, and S. Kollias, “Dense saliency-based spatiotemporal feature points for action recognition,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2009, IEEE, Los Alamitos.

[9] Y. Zhang and Q. Ji, “Active and dynamic information fusion for facial expression understanding from image sequences,” IEEE Trans. Pattern Anal. Mach.Intell May 2005, vol.27, no.5, pp.699-714.

[10] H. Gunes and M. Piccardi, “Bi-modal emotion recognition from expressive face and body gestures,” Journal of Network and Computer Applications, 2007.

[11] D.G. Lowe, “Distinctive image features from scale-invariant keypoints ,” IJCV, Nov 2004, vol.60, no.2, pp.91-110.

[12] W. Zheng, X. Zhou, C. Zou, L. Zhao, “Facial expression recognition using kernel canonical correlation analysis (KCCA),” IEEE Transactions on Neural Networks, 2006, Vol.17, No.1, pp.233-238.

[13] H. Gunes and M. Piccardi, “A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior,” In: Proc. Int. Conf. Pattern Recog, 2006, vol.1, pp.1148-1153.