IEICE TRANS. INF. & SYST., VOL.E97–D, NO.3 MARCH 2014

LETTER

Integrating Facial Expression and Body Gesture in Videos for Emotion Recognition

Jingjie YAN[†,††a)], **Wenming ZHENG**[††b)], ***Nonmembers***, **Minhai XIN**[††], ***Member***, ***and*** **Jingwei YAN**[††], ***Nonmember***

SUMMARY In this letter, we research the method of using face and gesture image sequences to deal with the video-based bimodal emotion recognition problem, in which both Harris plus cuboids spatio-temporal feature (HST) and sparse canonical correlation analysis (SCCA) fusion method are applied to this end. To efficaciously pick up the spatio-temporal features, we adopt the Harris 3D feature detector proposed by Laptev and Lindeberg to find the points from both face and gesture videos, and then apply the cuboids feature descriptor to extract the facial expression and gesture emotion features [1], [2]. To further extract the common emotion features from both facial expression feature set and gesture feature set, the SCCA method is applied and the extracted emotion features are used for the biomodal emotion classification, where the K-nearest neighbor classifier and the SVM classifier are respectively used for this purpose. We test this method on the biomodal face and body gesture (FABO) database and the experimental results demonstrate the better recognition accuracy compared with other methods.

key words: *bimodal emotion recognition, Harris plus cuboids spatio-temporal feature (HST), sparse canonical correlation analysis (SCCA)*

1. Introduction

The research of emotion recognition from multiple modalities such as speech, facial expression, gesture and electroencephalogram (EEG) has been an active research topic in affective computing, pattern recognition, artificial intelligence, and computer vision [3]–[5]. Just akin to facial expression recognition, vast majority of multimodal emotion recognition approaches focus on investigating the following two types of issues, i.e., the frame-based emotion recognition method and the sequence-based emotion recognition method [4], [6]–[9].

The frame-based multimodal emotion recognition methods usually regard those selected apex frames of the primordial video sequences as the representative emotion data to carry out the next feature extraction step and final recognition procedure [8]–[10]. For instance, Gunes et al. [9] present a frame-based approach to investigate the significance of temporal division or phases detection and division synchronization between facial expression frames and body gesture frames, and then they fuse the features gained from two modalities' apex frames by means of utilizing two fusion approaches. Comparing with those frame-based methods, sequence-based multimodal emotion recognition methods frequently adopt the entire video sequences as the fundamental emotion data for recognition [8], [9]. Petridis et al. [8] contrast the performance of the frame-based methods and the sequence-based methods for multimodal emotion recognition, and they employ Hidden Markov Model (HMM) and its other improvement forms to carry out the recognition of the entire video. Their results display that the property of two types of multimodal emotion recognition methods is probably approximate to some extent [8].

In the territory of video-based action recognition, local feature approach and globe feature approach are widespread utilized [11]. Among those local feature approaches, two most emblematic approaches are Laptev's spatio-temporal feature method [1] and Dollar's spatio-temporal feature [2] method, and which are on the basis of of Harris 3D and cuboid respectively [12]. In accordance with Dollar's spatio-temporal feature approach, shan et al. [6] recently discuss emotion recognition from single mode containing body gesture and facial expression, and then integrate the collected spatio-temporal feature of the forgoing two modalities by means of canonical correlation analysis.

In deal with emotion recognition, it is well known that the fusion of multiple modalities will achieve better recognition performance than using a single modality [5], [6], [13]–[15]. Nevertheless, it is very important to choose an effective fusion strategy in this work. A good fusion strategy will improve the recognition accuracy whereas a poor fusion strategy may not result in good performance, which has been testified by a number of works in the literatures [5], [6], [13]–[15]. For instance, Gunes and Piccardi [10], [13] independently conduct the emotion recognition problem from the single modality (facial expression or body gesture), and bimodal emotion recognition on account of the fusion form of the feature-level and the decision-level. Their test results reveal the ameliorating of recognition performance and robustness is received in the way of bimodal emotion recognition [10], [13].

The emotion information delivered by the facial expression modality and body gesture modality are rooted in the identical person and they are deemed to be relevant and practically concurrent to a certain degree [6], [9]. For example, the literature of [9] decomposes the original facial expression video and body gesture video as the form of tempo-

Manuscript received June 7, 2013.
Manuscript revised November 4, 2013.
[†]The author is with the School of Information Science and Engineering, Southeast University, Nanjing, 201196, China.
[††]The authors are with the Key Laboratory of Child Development and Learning Science (Ministry of Education), Research Center for Learning Science, Southeast University, Nanjing, 201196, China.
a) E-mail: yanjingjie1212@163.com
b) E-mail: wenming_zheng@seu.edu.cn
DOI: 10.1587/transinf.E97.D.610

ral dynamics (such as neutral, onset, apex, and offset), and its frame-by-frame contrastive result show that the temporal dynamics procedure of facial expression and body gesture are approximately synchronous [3].

In this letter, we research the method of using face and gesture image sequences to deal with the video-based bimodal emotion recognition problem, in which both Harris plus cuboids spatio-temporal feature (HST) and sparse canonical correlation analysis (SCCA) fusion method are applied to this end. We adopt the Harris 3D feature detector to find the points from both face and gesture videos, and then apply the cuboids feature descriptor to extract the facial expression and gesture emotion features [1], [2], [12]. To further extract the common emotion features from both facial expression feature set and gesture feature set, the SCCA method is applied and the extracted emotion features are used for the biomodal emotion classification.

The remaining of this literature is formed as the following. In Sect. 2, we give a introduction of the Harris plus cuboids spatio-temporal feature (HST) method. Feature fusion of facial expression and body gesture based on sparse canonical correlation analysis (SCCA) is introduced in Sect. 3. Eventually, certain tests are implemented on the FABO database in Sect. 4, and Sect. 5 concludes the letter and conducts some discussions.

2. Harris Plus Cuboids Spatio-Temporal Feature

In the next section, we will give a simple description of Harris plus cuboids spatio-temporal feature (HST) method. Just as the previous introduction, Laptev's spatio-temporal feature method [1] and Dollar's spatio-temporal feature [2] method are extensively applied in action recognition sphere.

Different from Dollar's spatio-temporal feature which in view of separable linear filters detector, the HST method employs the Harris 3D feature detector to substitute for separable linear filters detector [1], [2], [12]. In a general way, the HST method comprises of interest point detector, cuboids extraction, cuboids descriptor, cuboids library and cuboids cluster [1], [2], [6], [17], [19].

Suppose $b(x, y, t)$ represents the facial expression video data or body gesture video data, Laptev and Lindeberg define the following convolution operation formulation [1], [11], [19]

$$\mathbf{I}(x, y, t; \sigma^2, \tau^2) = g(x, y, t; \sigma^2, \tau^2) * b(x, y, t), \tag{1}$$

where σ and τ are spatial parameters and temporal parameters respectively, and function $g(x, y, t; \sigma^2, \tau^2)$ signifies a three-dimensional gaussian kernel function [1], [11], [19].

Then the authors again define a 3×3 three-dimensional second-moment matrix α as the following form [1], [11], [19]

$$\alpha = g(x, y, t; \sigma_i^2, \tau_i^2) * (\nabla \mathbf{I})(\nabla \mathbf{I})^T, \tag{2}$$

where $\nabla \mathbf{I} = (I_x, I_y, I_t)^T$ represents three-dimensional derivations with respect to x, y, t.

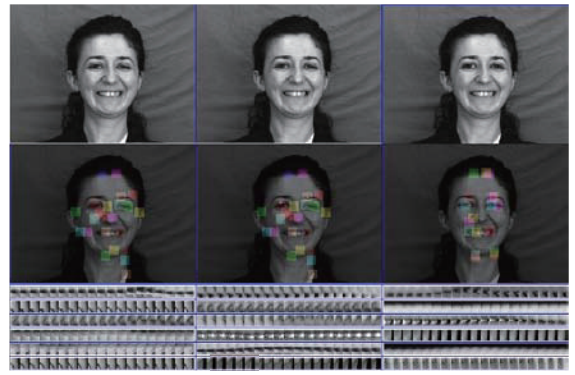

Fig. 1 Instances of the a number of cuboids picked from the FABO database using the HST method.

Ultimately, they present the following formulation to seek desirable interest point by calculating its maxima [1], [11], [19]

$$\begin{aligned} \mathbf{Q} &= \det(\alpha) - k(trace(\alpha))^3 \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3, \end{aligned} \tag{3}$$

where λ_1, λ_2, λ_3 are corresponding three eigenvalues and k is a constant.

After taking the above procedure, the form of cuboids feature descriptor is applied to convert and describe the movement information of initial video data similar to Dollar's spatio-temporal feature approach [2], [6], [11], [12].

Finally, the HST method also exploits the brightness gradient conversion method, cuboid to vector transformations approach and the form of histogram vector to stand for those achieved cuboids [2], [6], [11]. The detailed information please see the literature of [1] and [2]. Figure 1 shows some instances of cuboids picked from the FABO database using the HST method.

3. Integration of Facial Expression and Body Gesture Based on SCCA

In this section, we will review the sparse canonical correlation analysis (SCCA) method and then apply it to extract the emotional features from both facial expression feature set and gesture feature set obtained in aforementioned section [17], [18].

Let $\mathbf{F}$ and $\mathbf{B}$ denote the facial expression HST feature matrix and the body gesture HST featue matrix, respectively. Suppose that ϕ_F and ϕ_B are a pair of projection vectors such that the correlation between $\phi_F^T\mathbf{F}$ and $\phi_B^T\mathbf{B}$ is maximal, where the optimal projection vectors can be determined by the following optimization problem [6], [16], [21]:

$$\arg\max_{\phi_F, \phi_B} \frac{\phi_F^T \mathbf{K}_{FB} \phi_B}{\sqrt{\phi_F^T \mathbf{K}_{FF} \phi_F}\sqrt{\phi_B^T \mathbf{K}_{BB} \phi_B}}, \tag{4}$$

where $\mathbf{K}_{FB} = \mathbf{F}\mathbf{B}^T$, $\mathbf{K}_{FF} = \mathbf{F}\mathbf{F}^T$, and $\mathbf{K}_{BB} = \mathbf{B}\mathbf{B}^T$.

In solving the projection vectors ϕ_F and ϕ_B, Parkhomenko et al. [18] use a singular value decomposition

(SVD) [16]–[18], [22] solution approach. In this method, the SVD operator is first applied to **H**, such that

$$\mathbf{H} = \mathbf{K}_{FF}^{-1/2}\mathbf{K}_{FB}\mathbf{K}_{BB}^{-1/2} = \sum_{i=1}^{t} d_i x_i y_i^T. \quad (5)$$

Then, they expressed the SVD problem as a reduced-rank regression optimization problem and then imposed an ℓ_1 norm penalty on the principal SVD eigenvectors to achieve sparse SVD [22]. Based on the sparse SVD algorithm, they finally got the pair of sparse projection vectors ϕ_F and ϕ_B. Details of the SCCA method please see the the literature of [18].

By using the SCCA algorithm, we obtain two groups of sparse projection vectors $\phi_{F,i}$ and $\phi_{B,i}$ $(i = 1, \cdots, r)$, where r is the number of projection vector pairs. Then, we project the facial expression feature vectors and gesture feature vectors onto the projection vectors $\phi_{F,i}$ and $\phi_{B,i}$, respectively, and the projection results are finally concatenated to form a composed emotional feature vector for emotion classification [6].

4. Experiments

In this section, we implement a number of tests with single modality recognition and fusion recognition on account of the biomodal face and body gesture (FABO) database [5], [6], [8]–[10], [20], [23]. Four images of the FABO database are displayed in Fig. 2.

In our experiment, we choose a sub set of the FABO database which incorporates 113 facial expression videos and 113 body gesture videos. The emotion category of the foregoing selected videos incorporates boredom, disgust, happiness, puzzlement and uncertainty. Furthermore, each video's resolution is adjusted as 256×192. We regard PCA fusion method, CCA fusion method and the method of [6] as our contrast approach. Moreover, before conducting single modality recognition and fusion recognition, we project the extracted HST feature and Dollar's spatio-temporal (ST) feature of two modalities onto low dimension by means of PCA method. We compare certain different dimension and display the best result in our literature. We employ the "leave one subject out" cross-validation strategy and two types of classifiers incorporating the K-nearest neighbor classifier (euclidean distance) and the SVM classifier to recognize various emotion sorts.

Firstly, we implement the single modality recognition in view of the HST feature and the ST feature. Then the foregoing fusion method are adopted to integrate facial expression and body gesture by utilizing two classifiers. The single modality recognition rate (%) and four fusion methods recognition rate (%) are displayed in Table 1. In addition, the confusion matrices of SCCA+HST method under two classifiers are showed in Fig. 3 and Fig. 4

Fig. 2 Four images of the FABO database.

Table 1 The single modality recognition rate (%) and four fusion methods recognition rate (%) with the SVM classifier and the K-nearest neighbor classifier.

Method	the k-nearest neighbor classifier	the SVM classifier
ST (Body)	43.90	56.10
ST (Face)	48.78	54.47
HST (Body)	47.97	60.16
HST (Face)	47.15	58.54
HST+PCA	52.85	59.35
ST+CCA	52.85	62.60
HST+CCA	51.22	61.79
HST+SCCA	55.28	65.85

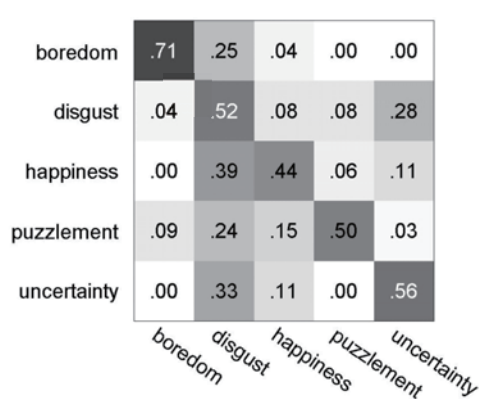

Fig. 3 The confusion matrice of HST+SCCA for bimodal emotion recognition using the K-nearest neighbor classifier.

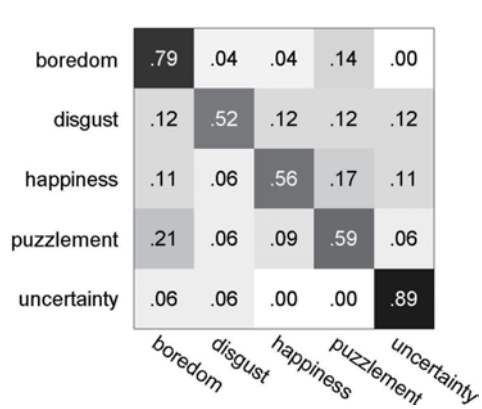

Fig. 4 The confusion matrice of HST+SCCA for bimodal emotion recognition using the SVM classifier.

On account of the foregoing results, we can note that the single modality recognition in view of the HST feature method is better then the ST feature method. The possible reason is the ST feature is appropriate for the case of large movement variation range [2], [6], but in our experiment, the each video's resolution is not high and the range of facial expression and body gesture is not sufficient strong. Consequently, those video data can not produce adequate interest points and the HST feature method based on the Harris 3D feature detector may be more appropriate for this situation compared to the ST feature method.

In addition, from Table 1, we can observe that four fusion methods are mainly better than single modality recognition. Besides, the SCCA method receives more higher recognition rate with respect to those contrast approaches under two classifiers. This is probable because SCCA method is capable of effectively learning the emotion information from both facial expression and gesture feature set. In addition, the memory size of HST+SCCA method and ST+CCA method are about $300\,M$ to $1.0\,G$ and $400\,M$ to $1.1\,G$ respectively.

5. Conclusions and Discussion

In this letter, we research video-based bimodal emotion recognition based on Harris plus cuboids spatio-temporal feature (HST) and sparse canonical correlation analysis (SCCA) fusion method. A number of tests on the biomodal face and body gesture (FABO) database demonstrate the better recognition accuracy compared to other relevant methods. However, the method also has one major shortcoming that it can only deal with two modalities such as facial expressions and body gestures. To work out this shortcoming, we may can utilize multiset canonical correlations analysis (MCCA) [24] to integrate three or more modality such as facial expressions, body gestures, speech and so on.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61231002 and Grant 61073137, the Natural Science Foundation of Jiangsu Province under Grant BK20130020, and the Program for Distinguished Talents of Six Domains in Jiangsu Province of China under Grant 2010-DZ088.

References

[1] I. Laptev and T. Lindeberg, "Space-time interest points," ICCV, pp.432–439, 2003.
[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatiotemporal features," VS-PETS, pp.65–72, 2005.
[3] S. Chen, Y. Tian, Q. Liu, and D.N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," IEEE Int'l Conf. Computer Vision and Pattern Recognition workshop for Human Communicative Behavior Analysis, pp.7–12, 2011.
[4] C. Shan and R. Braspenningn, Part IV. Recognizing Facial Expressions Automatically from Video, Handbook of Ambient Intelligence and Smart Environments, Springer, 2010.
[5] H. Gunes, M. Piccardi, and M. Pantic, "Affective computing: focus on emotion expression, synthesis, and recognition," Austria: InTech Education and Publishing, chap 10. From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities, 2008.
[6] C. Shan, S. Gong, and P.W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," Proc. Brit. Mach. Vis. Conf., pp.1–10, 2007.
[7] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Simultaneous estimation of facial pose and expression by combining particle filter with gradient method," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J92-D, no.8, pp.1349–1362, Aug. 2009.
[8] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. Dynamic Modeling of Human Nonverbal Behavior from Multiple Cues and Modalities," ICMI-MLMI, pp.23–30, 2009.
[9] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," IEEE Trans. Syst. Man. Cybern.-Part B, vol.39, no.1, pp.64–84, 2009.
[10] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," J. Netw. Comput. Applicat., vol.30, no.4, pp.1334–1345, 2007.
[11] N.M. Nayak, R.J. Sethi, B. Song, and A.K. Roy-Chowdhury, Modeling and recognition of complex human activities, Visual Analysis of Humans, chap 15, Springer, 2011.
[12] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid1, "Evaluation of local spatio-temporal features for action recognition," BMVC, 2009.
[13] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," Proc. IEEE SMC, vol.4, pp.3437–3443, 2005.
[14] C. Busso, Z. Deng, S. Yildirim, M. Bulut, and C.M. Lee, "Analysis of emotion recognition using facial expressions, speech and multimodal information," ICML, 2004.
[15] Z. Zeng, M. Pantic, and T.S. Huang, Emotion recognition based on multimodal information, Affective Information Processing, chap 14, Springer, 2009.
[16] X. Zhou, W. Zheng, and M. Xin, "Improving CCA via spectral components selection for facial expression recognition," IEEE ISCAS, pp.1696–1699, 2012.
[17] W. Qiu, A Study on the Methods of Biomal Emotion Recognition based on Body Gesture and Facial Expression, master thesis, Southeast University, 2011. (in Chinese).
[18] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," Statistical Applications in Genetics and Molecular Biology, vol.8, pp.55–61, 2009.
[19] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," ICPR, pp.32–36, 2004.
[20] J. Yan, W. Zheng, M. Xin, and W. Qiu, "Bimodal emotion recognition based on body gesture and facial expression," J. Image and Graphics, vol.18, pp.1101–1106, 2013. (in Chinese)
[21] Y.H. Yuan, Q.S. Sun, Q. Zhou, and D.S. Xia, "A novel multiset integrated canonical correlation analysis framework and its application in feature fusion," Pattern Recognit., vol.44, pp.1031–1040, 2011.
[22] H. Shen and J. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," J. Multivariate Analy., vol.99, pp.1015–1034, 2008.
[23] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," ICPR, pp.1148–1153, 2006.
[24] A.A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," IEEE Trans. Image Process., vol.11, no.3, pp.293–305, 2002.