MIC-TJU in MediaEval 2017 Emotional Impact of Movies Task

Yun Yi[1,2], Hanli Wang[2,*], Jiangchuan Wei[2]
[1]Department of Mathematics and Computer Science, Gannan Normal University, Ganzhou 341000, China
[2]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

ABSTRACT

To predict the emotional impact and fear of movies, we propose a framework which employs four audio-visual features. In particular, we utilize the features extracted by the methods of motion keypoint trajectory and convolutional neural networks to depict the visual information, and extract a global and a local audio features to describe the audio cues. The early fusion strategy is employed to combine the vectors of these features. Then, the linear support vector regression and support vector machine are used to learn the affective models. The experimental results show that the combination of these features obtains promising performances.

1 INTRODUCTION

The 2017 emotional impact of movies task is a challenging task, which contains two subtasks (*i.e.*, valence-arousal prediction and fear prediction). A brief introduction about this challenge has been given in [3]. In this paper, we mainly introduce the system architecture and algorithms used in our framework, and discuss the evaluation results.

2 FRAMEWORK

The key components of the proposed framework is shown in Fig. 1, and the highlights of our framework are introduced below.

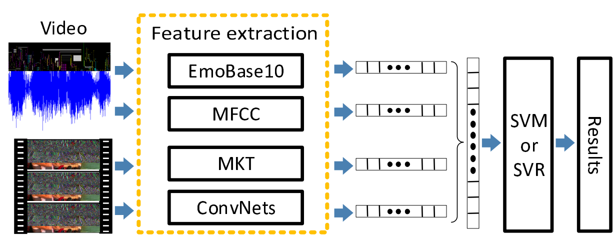

Figure 1: An overview of the key components of the proposed framework.

2.1 Feature Extraction

In this framework, we evaluate four features, including EmoBase10 feature [5], Mel-Frequency Cepstral Coefficients (MFCC) feature [4], Motion Keypoint Trajectory (MKT) feature [15], and Convolutional Networks (ConvNets) feature [12, 14].

2.1.1 MFCC Feature. In affective content analysis, audio modality is essential. MFCC is a famous local audio feature. The time window of MFCC is set to 32 ms, and set 50% overlap between two adjacent windows. In order to promote the performance, we append delta and double-delta of 20-dimensional vectors into the original MFCC vector. Therefore, a 60-dimensional MFCC vector is generated. We apply Principal Component Analysis (PCA) to reduce the dimension of the local feature, and use the Fisher Vector (FV) model [10] to represent a whole audio file via a signature vector. The cluster number of Gaussian Mixture Model (GMM) is set to 512, and the signed square root and L2 norm are utilized to normalize the vectors. In our experiments, we use the toolbox provided by [4] to calculate the vectors of MFCC.

2.1.2 EmoBase10 Feature. To depict audio information, we extract the EmoBase10 feature [5, 11], which is a global and high-level audio feature. As suggested by [5, 11], the default parameters are utilized to extract the 1,582-dimensional vector of EmoBase10. The 1,582-dimensional vector results from: (1) 21 functionals applied to 34 Low-Level Descriptors (LLD) and 34 corresponding delta coefficients, (2) 19 functionals applied to the 4 pitch-based LLD and their 4 delta coefficient contours, (3) the number of pitch onsets and the total duration of the input [5, 11]. Then, the signed square root and L2 norm are utilized to normalize the vectors. We calculate the EmoBase10 feature by using the openSMILE[1] toolkit.

2.1.3 MKT Feature. We utilize the MKT [15] Feature to depict the motion information. Motion keypoints are tracked by the approach of MKT at multiple spatial scales, and an optical flow rectification algorithm that is based on vector field consensus [9] is designed to reduce the influence of camera motions. To depict trajectories in a video, we calculate four local descriptors along trajectories, including Histogram of Oriented Gradient (HOG) [1], Motion Boundary Histogram (MBH) [2], Histogram of Optical Flow (HOF) [8] and Trajectory-Based Covariance (TBC) [15]. In general, MBH and HOF represent the local motion information, HOG

*Hanli Wang is the corresponding author (hanliwang@tongji.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61622115 and Grant 61472281, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005).

MediaEval'17, 13-15 September 2017, Dublin, Ireland

[1]http://audeering.com/technology/opensmile

describes the local appearance, and TBC depicts the relationships between di erent motion variables. After calculating these local vectors, we individually apply the RootSIFT normalization (*i.e.*, square root on each dimension after L1 normalization) to normalize these vectors.

In order to reduce the dimension of descriptors, we apply PCA to the four descriptors individually. Then, the FV model [10] is used to encode these local vectors. In particular, we apply GMM to construct a codebook of each descriptor, and set the number of GMM to 128. Finally, the signed square root and L2 normalization are applied to these vectors. To combine the trajectory-based descriptors, we concatenate the vectors of these four descriptors into a single one.

2.1.4 ConvNets Feature. Convolutional Neural Networks (CNNs) have been successfully applied in many areas. The two-stream Convolutional Networks (ConvNets) feature include two streams [12, 14], *i.e.*, the spatial stream ConvNet and temporal stream ConvNet. The spatial ConvNet operating on video frames indicates the information about scenes and objects. Meanwhile, the temporal ConvNet stacking optical ow elds conveys the motion information of videos. The two-stream ConvNets feature is calculated according to the processes in [12, 14] based on the network architecture of BN-Inception [7].

In our experiments, the Ca e toolbox is used to calculate the ConvNets feature. We utilize the models pretrained on the UCF101 dataset [13], and calculate the feature vectors from the `global_pool' layer. Let the sets of vectors extracted from spatial and temporal nets be individually denoted as $\mathbb{S} = \{S_1, \cdots, S_i, \cdots, S_N\}$ and $\mathbb{T} = \{T_1, \cdots, T_i, \cdots, T_N\}$, where N is the number of frames, and S_i and T_i are 1,024-dimensional vectors. To depict a video via one vector, we utilize two strategies, including Fisher Vector (FV) and Mean Standard Deviation (MSD). The feature vectors calculated by the two strategies are denoted as ConvNets-FV and ConvNets-MSD separately. For the extraction of ConvNets-FV, we follow the processes as suggested in [10, 15, 16], and set the cluster number of GMM to 64. For the feature calculation of ConvNets-MSD, we calculate the mean of the two sets respectively, which are denoted as $\mu(\mathbb{S})$ and $\mu(\mathbb{T})$, and calculate their standard deviations denoted as $\sigma(\mathbb{S})$ and $\sigma(\mathbb{T})$. Then, the four vectors (*i.e.*, $\mu(\mathbb{S})$, $\mu(\mathbb{T})$, $\sigma(\mathbb{S})$, and $\sigma(\mathbb{T})$) are concatenated to produce a $(1,024 \times 4)$-dimensional vector.

2.2 Regression and Classification

In the two subtasks, we employ linear Support Vector Regression (SVR) and Support Vector Machine (SVM) [6] to learn the emotional models separately. For the fear subtask, the number of positive samples is less than that of the negative samples. To solve this problem, we weight positive and negative samples in an inverse manner. The regularization parameter C is set by cross-validation on the training set. The LIBLINEAR toolbox[2] is utilized to implement the L2-regularized L2-loss SVM and SVR.

3 RESULTS AND DISCUSSIONS

In this task, we submit 5 runs, and the results are given in Table 1 and Table 2. The main di erence of these 5 runs is the selection of features. We select MFCC, ConvNets-MSD and EmoBase10 in Run 1, MFCC and ConvNets-MSD in Run 2, MFCC, ConvNets-FV and EmoBase10 in Run 3, MFCC, ConvNets-MSD, EmoBase10 and MKT in Run 4, and MFCC, ConvNets-FV, EmoBase10 and MKT in Run 5. For the valence-arousal subtask, we report Mean Square Error (MSE) and Pearson Correlation Coe cient (PCC) [3]. For the fear subtask, the performances of accuracy, precision, recall and F1-score are considered as suggested in [3]. Regarding the learning processes of all runs, we utilize SVR in the valence-arousal subtask, and use SVM in the fear subtask.

Table 1: Results of the valence-arousal subtask.

Runs	Valence		Arousal	
	MSE	PCC	MSE	PCC
Run 1	0.21972	0.10818	0.15119	-0.02392
Run 2	0.21756	0.11622	0.15236	-0.03570
Run 3	**0.21271**	**0.1533**	0.13989	**0.08182**
Run 4	0.22661	0.09801	**0.12812**	-0.01139
Run 5	0.22090	0.07849	0.13472	0.05013

Table 2: Results of the fear subtask.

Runs	Accuracy	Precision	Recall	F1-score
Run 1	**0.862307**	**0.375595**	0.099091	0.142365
Run 2	0.848925	0.368764	0.072547	0.096831
Run 3	0.840726	0.114286	0.023183	0.038265
Run 4	0.845466	0.171429	0.029288	0.039684
Run 5	0.844685	0.214286	0.016592	0.029383

As shown in Table 1 and Table 2, Run 3 obtains the best result in the valence-arousal subtask, and Run 1 achieves the top performance in the fear subtask. This partly demonstrates that more features do not necessarily achieve better result and di erent combinations of features are suitable for di erent subtasks. By comparing the results of Run 1 and Run 3, we can nd that ConvNets-FV is suitable for the valence-arousal subtask and ConvNets-MSD is suitable to depict fear.

REFERENCES

[1] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR'05*. 886{893.

[2] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human detection using oriented histograms of ow and appearance. In *ECCV'06*. 428{441.

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/multicore-liblinear

[3] Emmanuel Dellandrea, Martijn Huigsloot, Liming Chen, Yoann Baveye, and Mats Sjöberg. 2017. The MediaEval 2017 Emotional Impact of Movies Task. In *MediaEval 2017 Workshop*.

[4] Daniel P. W. Ellis. 2005. PLP and RASTA (and MFCC, and inversion) in Matlab. http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/. (2005). online web resource.

[5] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM MM'13*. 835{838.

[6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classi cation. *Journal of Machine Learning Research* 9 (2008), 1871{1874.

[7] Sergey Io e and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML'15*. 448{456.

[8] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *CVPR'08*. 1{8.

[9] Jiayi Ma, Ji Zhao, Jinwen Tian, Alan L Yuille, and Zhuowen Tu. 2014. Robust point matching via vector eld consensus. *IEEE Trans. Image Processing* 23, 4 (2014), 1706{1721.

[10] Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *CVPR'07*.

[11] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Interspeech'10*.

[12] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS'14*. 568{576.

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01* (2012).

[14] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: towards good practices for deep action recognition. In *ECCV'16*. 20{36.

[15] Yun Yi and Hanli Wang. 2017. Motion keypoint trajectory and covariance descriptor for human action recognition. *The Visual Computer* (2017), 1{13.

[16] Yun Yi, Hanli Wang, and Bowen Zhang. 2017. Learning correlations for human action recognition in videos. *Multimedia Tools and Applications* 76, 18 (2017), 18891{18913.