RUC at MediaEval 2016 Emotional Impact of Movies Task: Fusion of Multimodal Features

Shizhe Chen, Qin Jin
School of Information, Renmin University of China, China
{cszhe1, qjin}@ruc.edu.cn

ABSTRACT

In this paper, we present our approaches for the Mediaeval Emotional Impact of Movies Task. We extract features from multiple modalities including audio, image and motion modalities. SVR and Random Forest are used as our regression models and late fusion is applied to fuse different modalities. Experimental results show that the multimodal late fusion is beneficial to predict global affects and continuous arousal and using CNN features can further boost the performance. But for continuous valence prediction the acoustic features are superior to other features.

1. INTRODUCTION

The 2016 Emotion Impact of Movies Task [1] involves two subtasks: global and continuous affects prediction. The global subtask requires participants to predict the induced valance and arousal values for the short video clips, while the affects values should be continuously predicted every second for long movies in the continuous subtask. In the following sections, we describe the multimodal features, models and experiments in details.

2. FEATURE EXTRACTION

2.1 Audio Modality

Statistical Acoustic Features: Statistical acoustic features are proved to be very effective in speech emotion recognition. We use the open-source toolkit OpenSMILE [2] to extract three kinds of features `IS09`, `IS10` and `IS13`, which uses the configuration in INTERSPEECH 2009 [3], 2010 [4] and 2013 [5] Paralinguistic challenge respectively. The difference between these features is that features in the later years cover more low-level features and statistical functions.

MFCC-based Features: The Mel-Frequency Cepstral Coefficients (MFCCs) [6] are the most widely used low-level features. Therefore, we use MFCCs as our frame-level feature and apply two encoding strategies, Bag-of-Audio-Words (BoW) [7] and Fisher Vector Encoding (FV) [8], to transform the set of MFCCs to the sentence-level features. For `mfccBoW` features, the acoustic codebook is trained by K-means with 1024 clusters. For `mfccFV` features, we use the GMM to train the codebook with 8 mixtures.

In the continuous subtask, the audio features are extracted with the window of 10s and shift of 1s to cover more context.

2.2 Image Modality

Hand-crafted Visual Features: We extract the Hue-Saturation Histogram (`hsh`) to describe the color information and the Dense SIFT (`DSIFT`) features to represent the visual appearance information. For `hsh` features, we quantize the hue to 30 levels and the saturation to 32 levels. For `DSIFT` features, we use Fisher Vector encoding to construct the video-level features. Then kernel PCA is utilized to reduce the dimensionality into 4096.

DCNN Features: To explore the performance of different pre-trained CNN models, we extract multiple layers from different CNN models including inception-v3 [9], VGG-16 and VGG-19 [10]. All the CNN features are applied with mean pooling to generate video-level representations.

2.3 Motion Modality

To exploit the temporal information in the video, we extract the improved Dense Trajectory (iDT) [11] and the C3D features [12].For iDT features, HOG, HOF and MBH features are densely extracted from the video and encoded with Fisher Vector. Then kernel PCA is used to reduce dimensionality into 4096. For C3D features, we extract activations from the penultimate layer for every non-overlap 16 frames and use mean pooling to aggregate them into one vector.

The challenge also provides baseline features [13] for global subtask, which consists of acoustic and visual features.

3. EXPERIMENTS

3.1 Experimental Setting

In the global subtask, there are 9,800 video clips from 160 movies in the development set. We randomly select 6093, 1761 and 1946 videos as our local training, validation and testing sets respectively. Video clips from the same movies are kept in the same set. In the continuous subtask, the 30 movies in the development set are also divided into 3 parts with 24 for training, 3 for validation and 3 for testing.

We train SVR and Random Forest for each kind of features and use grid search to select the best hyper-parameters. For SVR, we explore linear and RBF kernels and tune the cost from 2^{-5} to 2^{12} and the epsilon-tube from 0.1 to 0.4. For Random Forest, the number of trees and the depth of trees are tuned from 100 to 1000 and from 3 to 20 respectively. We apply late fusion to fuse different features by training a second-layer model (linear SVR) with input of the best predictions for each kind of features using the local validation set. We use Sequential Backward Selection algorithm to find the best subset of feature types for late fusion.

Copyright is held by the author/owner(s).
MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Figure 1: MSE Performance of Different Features for Global Arousal Prediction on Local Testing Set

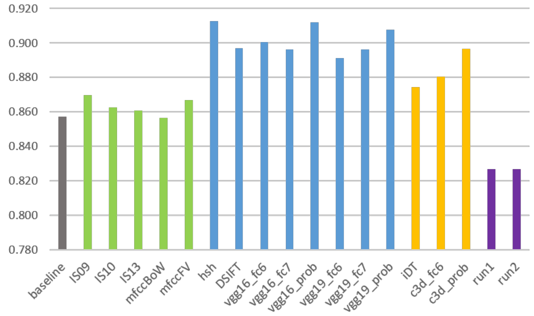
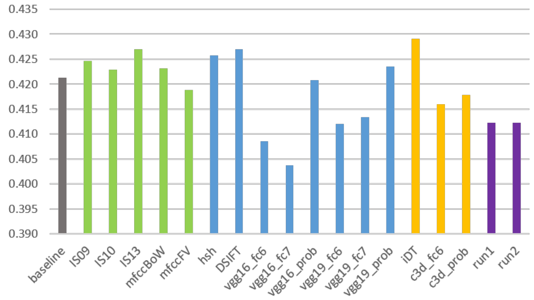


Figure 2: MSE Performance of Different Features for Global Valence Prediction on Local Testing Set



3.2 Global Affects Prediction

In the global subtask, we use the mean standard error (MSE) as evaluation metric. Figure 1 presents MSE of the different features for arousal prediction. The audio modality performs the best. Since the baseline feature contains multimodal cues, it achieves the second best performance following our `mfccBoW` feature. The run1 is the late fusion of all the audio features, `baseline` and `iDT` features. In the run2 system, besides the features used in run1, `c3d_fc6`, `vgg16_fc7` and `vgg19_fc6` features are also used in late fusion. The arousal prediction performance is significantly improved by the multimodal late fusion.

The MSE of different features for global valence prediction is shown in Figure 2. The image modality features especially the CNN features are better than other modalities for valence prediction. The run1 system consists of `baseline`, `IS10`, `mfccBoW`, `mfccFv` and `hsh`. The run2 system also uses `c3d_fc6`, `c3d_prob`, `vgg16_fc6`, `vgg16_fc7` and the features in run1. Although the late fusion performance does not outperform the unimodal performance with CNN `vgg16` features on our local testing set, it might be more robust than using one single feature.

3.3 Continuous Affects Prediction

In the continuous subtask, we use the Pearson Correlation Coefficient (PCC) for performance evaluation instead of MSE. Because the labels in the continuous subtask have closer temporal connections than those in the global subtask and thus the shape of the prediction curve is more important. Since the testing set is relative small and the performance is quite unstable in the evaluation, we average the performance of the validation and testing set. Figure 3 shows the PCC results of different features. The `mfccFV` feature performs the best in both arousal and valence prediction. The settings for the submitted three runs are as follows. In run1, we apply late fusion over `mfccFV`, `IS09` and `IS10` for arousal and use the `mfccFV` SVR for valence. In run2, `mfccFV`, `IS09`, `IS10` and `inc_fc` are late fused for arousal and `mfccFV` and `IS09` are late fused for valence. The run3 late fuses `mfccFV`, `IS09` and `inc_fc` for arousal and use `mfccFV` Random Forest for valence. In our experiment, late fusion is beneficial for the arousal prediction but not for valence prediction.

Figure 3: PCC Performance of Different Features for Continuous Prediction on Local Testing Set

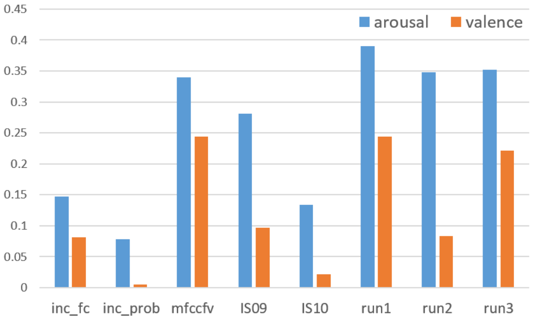


Table 1: The Submission Results for Global and Continuous Affects Prediction

		Arousal		Valence	
		MSE	PCC	MSE	PCC
Global	run1	**1.479**	0.405	0.218	0.312
	run2	1.510	**0.467**	**0.201**	**0.419**
Continuous	run1	**0.120**	0.147	0.102	0.106
	run2	0.121	**0.236**	0.108	0.132
	run3	0.122	0.191	**0.099**	**0.142**

3.4 Submitted Runs

In Table 1, we list our results on the challenge testing set. For the global subtask, comparing between run1 and run2, fusing CNN features can greatly improve the arousal and valence prediction performance. For the continuous subtask, the fusion of image and audio cues improves the arousal prediction performance. But for valence prediction, the `mfccFV` feature alone achieves the best results.

4. CONCLUSIONS

In this paper, we present the multimodal approach to predict global and continuous affects. The best result on the global subtask is achieved by the late fusion of audio, image and motion modalities. However, for the continuous subtask, the `mfccFV` feature significantly outperforms other features and benefits little from late fusion on valence prediction. In the future work, we will explore more features for continuous affects prediction and use LSTMs to model the temporal structure of the videos.

5. ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202.

6. REFERENCES

[1] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Sjöberg, and Christel Chamaret. The mediaeval 2016 emotional impact of movies task. In *MediaEval 2016 Workshop, Hilversum, Netherlands, Oct. 20-21*, 2016.

[2] Florian Eyben, Martin Llmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia, Mm*, pages 1459–1462, 2010.

[3] Björn W. Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pages 312–315, 2009.

[4] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011.

[5] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, and Erik Marchi. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings of Interspeech*, pages 148–152, 2013.

[6] Steven B. Davis. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in Speech Recognition*, 28(4):65–74, 1990.

[7] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1370–1374, 2014.

[8] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[11] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.

[13] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *ACII*, pages 77–83, 2015.