RUC at MediaEval 2016: Predicting Media Interestingness Task

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

ABSTRACT

Measuring media interestingness has a wide range of applications such as video recommendation. This paper presents our approach in the MediaEval 2016 Predicting Media Interestingness Task. There are two subtasks: image interestingness prediction and video interestingness prediction. For both subtasks, we utilize hand-crafted features and C-NN features as our visual features. For the video subtask, we also extract acoustic features including MFCC Fisher Vector and statistical acoustic features. We train SVM and Random Forest as classifiers and early fusion is applied to combine different features. Experimental results show that combining semantic-level and low-level visual features are beneficial for image interestingness prediction. When predicting video interestingness, the audio modality has superior performance and the early fusion of visual and audio modalities can further boost the performance.

1. SYSTEM DESCRIPTION

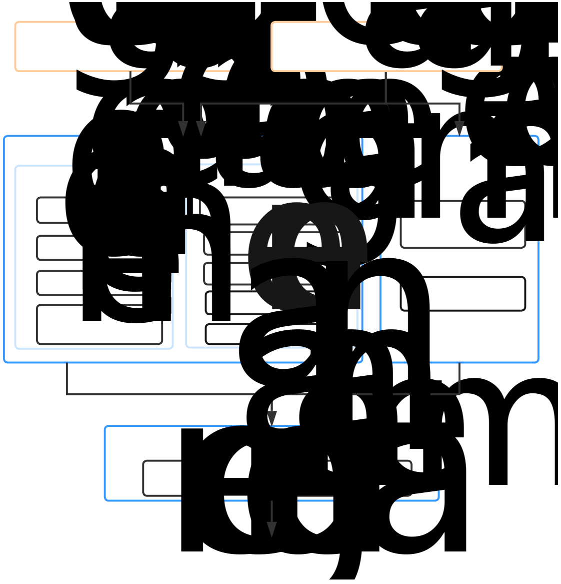
An overview of our framework in the MediaEval 2016 Predicting Media Interestingness Task [1] is shown in Figure 1. For image interestingness prediction, we use hand-crafted visual features and CNN features. For the video subtask, we utilize both visual and audio cues in the video to predict the interestingness. Early fusion is applied to combine different features. In the following subsections, we describe the feature representation and prediction model in details.

1.1 Feature Extraction

1.1.1 Visual Features

DCNN is the state-of-the-art model in many visual tasks such as object detection, scene recognition etc. In this task, we extract activations from the penultimate and the last softmax layers from the AlexNet and Inception-v3 [2] pretrained on ImageNet as our image-level CNN features, namely `alex_fc7`, `alex_prob`, `inc_fc`, `inc_prob` respectively. The features extracted from the last layers are the probability distribution on 1000 different objects, which describe the semantic level of concepts people might show interest in. The penultimate layer features are the abstraction of the image content and have shown great generalization ability in different tasks. We also use hand-crafted visual features including Color Histogram, GIST, LBP, HOG, Dense SIFT provided in [3] to cover different aspects of the images. For the video subtask, mean pooling is applied over all the image features of the video clip to generate video-level features.

Figure 1: An Overview of the System Framework



1.1.2 Acoustic Feature

Statistical Acoustic Features: Statistical acoustic features are proved to be effective in speech emotion recognition. We use the open-source toolkit OpenSMILE [4] to extract the statistical acoustic features, which use the configuration in INTERSPEECH 2009 [5] Paralinguistic challenge. Low-level acoustic features such as energy, pitch, jitter and shimmer are first extracted over a short-time window. And then statistical functions like mean, max are applied over the set of low-level features to generate sentence-level features.

MFCC based Features: The Mel-Frequency Cepstral Coefficients (MFCCs) [6] are the most widely used low-level features which have been successfully applied in many speech tasks. Therefore, we use MFCCs as our frame-level feature with window of 25ms and shift of 10ms. The Fisher Vector Encoding (FV) [7] is applied to transform the variant length of MFCCs to the sentence-level features. We train a Gaussian Mixture Models (GMMs) with 8 mixtures as our audio word dictionary. Then we compute the gradient of the log likelihood with respect to the parameters of the GMMs for

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Figure 2: MAP of Single Feature for Image Subtask on Local Testing Set

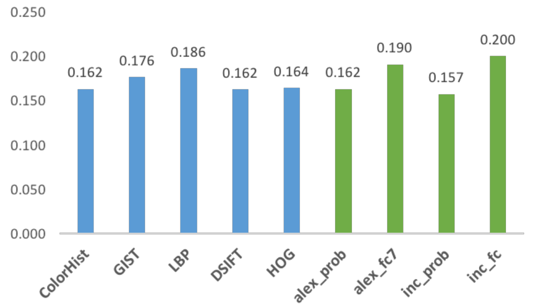


Figure 3: MAP of Early Fusion for Image Subtask on Local Testing Set

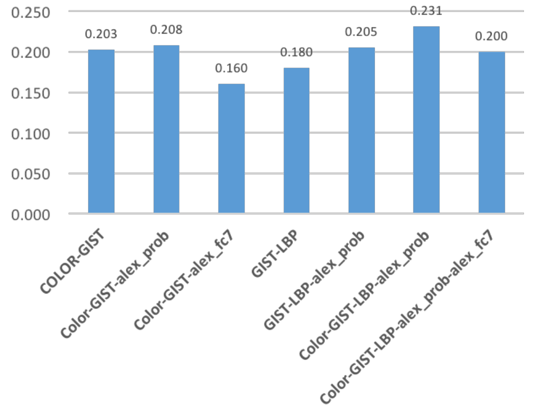


Figure 4: MAP of Single Feature for Video Subtask on Local Testing Set

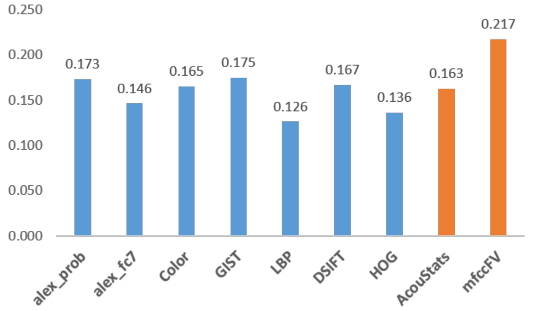


Table 1: MAP of Early Fusion for Image and Video Subtask on the Real Testing Set (the Official Evaluation Metric)

	features	model	real tst
image subtask	GIST-LBP-alex_prob	RF	0.199
	Color-GIST-alex_prob	RF	0.204
	Color-GIST-LBP-alex_prob	SVM	0.199
video subtask	AcouStats-GIST	SVM	0.165
	mfccFV-GIST	SVM	0.170

each audio to maximize the probability that the model can fit the data. L2-norm is applied for the mfccFV features.

1.2 Classification Model

For both the image and video systems, we train binary SVM and Random Forest as our interestingness classification models. Hyper parameters of the models are selected according to the mean average precision (MAP) on our local validation set using grid search. For SVM, RBF kernel is applied and the cost is searched from 2^{-2} to 2^{10}. And for Random Forest, the number of trees is set to be 100 and the depth of the tree is searched from 2 to 16.

2. EXPERIMENTS

2.1 Experimental Setting

There are 5054 images or videos in total for development in each subtask. We use video with id from 0 to 40 (4014 samples) as the local training set, 41 to 45 (468 samples) as local validation set and the remained videos (572 samples) as the local testing set. We use the whole development set to train the final submitted systems.

2.2 Experimental Results

Figure 2 shows the best MAP performance of SVM and Random Forest classifiers for each kind of features in the image subtask. The penultimate CNN features inc_fc and alex_fc7 achieve the top performance among all the visual features. However, the probability features extracted from CNN do not perform well alone.

We then use early fusion to concatenate different visual features. Figure 3 shows some of the fusion results. We can see that combining the alex_prob with other visual appearance features can significantly improve the classification performance, which shows that the semantic-level features and low-level appearance features are complementary. However, concatenating alex_fc7 with hand-crafted features do not bring any improvement.

For video interestingness prediction, Figure 4 presents the performance of each single feature. The audio modality outperforms the visual modality and mfccFV achieves the best performance. Fusing acoustic features with the best visual feature GIST are beneficial, for example, AcouStats-GIST achieves MAP of 20.80%, which is 19% relative gain compared with the MAP of single feature GIST.

The total five runs we submitted are listed in Table 1.

3. CONCLUSIONS

Our results show that image interestingness prediction can benefit from combining semantic-level objects probabilities distribution features and low-level visual appearance features. For predicting video interestingness, audio modality shows superior performance than visual modality and the early fusion of two modalities can further boost the performance. In the future work, we will explore ranking models for the interestingness prediction task and extract more discriminative features such as video motion features.

4. ACKNOWLEDGMENTS

This research was supported by the Research Funds of Renmin University of China (No. 14XNLQ01) and the Beijing Natural Science Foundation (No. 4142029).

5. REFERENCES

[1] C.-H. Demarty, M. Sj●berg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingness task. *In Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands, Oct. 20-21*, 2016.

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

[3] Yu-Gang Jiang, Qi Dai, Tao Mei, Yong Rui, and Shih-Fu Chang. Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, 17(8):1174{1186, 2015.

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

[5] Bj●rn 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.

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