MIC-TJU in MediaEval 2015 Affective Impact of Movies Task

Yun Yi Hanli Wang * Bowen Zhang Jian Yu

Department of Computer Science and Technology, Tongji University, Shanghai 201804, P. R. China
{13yiyun,hanliwang,102310,yujian}@tongji.edu.cn

ABSTRACT

The MediaEval 2015 Affective Impact of Movies task challenged participants to automatically detect video content that depicts violence, or predict the affective impact that video content will have on viewers. In this paper, we describe our system and discuss the performance results obtained in this task. We adopt our recently proposed Trajectory Based Covariance (TBC) descriptor to depict the motion information. Besides that, other features including audio, scene, color and appearance are also utilized in our system. To combine these features, a late fusion strategy is employed. Our results show that the trajectory based motion feature can achieve very competitive performances, furthermore the combination with audio, scene, color and appearance features can improve the overall performance.

1. INTRODUCTION

The Affective Impact of Movies task is a challenging task which requires to build a high performance system to automatically detect video content that depicts violence, or predict the affective impact that video content will have on viewers. This task contains two subtasks: Induced Affect Detection and Violence Detection. A brief introduction to the dataset for training and testing as well as evaluation metrics of these two subtasks has been given in [4]. In this paper, we mainly discuss the techniques and algorithms employed by our system, as well as the related system architecture and evaluation results.

2. SYSTEM DESCRIPTION

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

2.1 Feature Extraction

In the feature extraction part, five kinds of features are used including Mel-Frequency Cepstral Coefficients (MFCC) feature, Improved Dense Trajectory (IDT) based feature, Dense Scale Invariant Feature Transform (Dense SIFT) feature, Hue-Saturation Histogram (HSH) feature and Convolutional Neural Network (CNN) based feature.

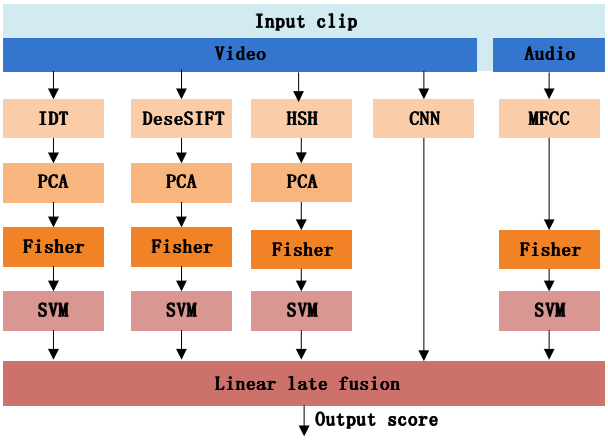

Figure 1: An overview of the key components.

2.1.1 *MFCC Feature*

We adopt the famous MFCC algorithm in the audio section [2]. The time window for MFCC is 32 ms and there is 50% overlap between two adjacent windows. To fully utilize the discrimination ability of MFCC, we append delta and double-delta of 20-dimension MFCC vectors into the original MFCC vector to generate a 60-dimension MFCC vector. To represent a whole audio file as a single vector, we adopt the classic Bag-of-Words (BoW) framework, where Fisher Vector and Gaussian Mixture Model (GMM) are used [3]. The cluster number of GMM is set to 512 in our system.

2.1.2 *IDT Based Feature*

The Improved Dense Trajectory (IDT) approach [5] is an efficient method to track human actions. The trajectory based descriptors, including the Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH), are employed in our system to depict the motion information of video content. In addition, our recently proposed TBC [6] descriptor is also utilized.

After the extraction of descriptors, these feature vectors are normalized with the $L1$ and signed square root normal-

*H. Wang is the corresponding author.

This work was supported in part by the "Shu Guang" project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23 and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005).

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

ization. To reduce the dimension of descriptors, the Principal Component Analysis (PCA) is individually applied to the three descriptors (*i.e.*, HOG, HOF, MBH) as in [5], and the Logarithm Principal Component Analysis (LogPCA) is applied to the TBC descriptor.

To encode feature vectors, the Fisher Vector model [3] is utilized. Specifically, GMM is applied to construct a codebook for each descriptor and we compute one Fisher Vector over an entire video followed by applying the signed square root and $L2$ normalization, which is able to significantly improve the performance in combination with linear SVM. To combine the IDT based descriptors, early fusion is performed to generate the final feature vector by concatenating the aforementioned four feature vectors into a single one.

2.1.3 Dense SIFT Feature

Scene information is an important cue for video content analysis. The Dense SIFT approach is utilized to depict scene information of video clips. We densely compute SIFT descriptors at multiple scales on a dense grid every 30 frames. After the SIFT descriptors are extracted, PCA is utilized to reduce the dimension of SIFT descriptors, and GMM is applied to construct a codebook. Unlike IDT based descriptors, we compute one Fisher Vector over one frame, and then average each Fisher Vector at the current temporal scale. In our system, we set the cluster number of GMM to 512 and the temporal scale number to 2.

2.1.4 Hue-Saturation Histogram Feature

The color of videos can affect the viewer's psychology, so the Hue-Saturation Histogram (HSH) is used to describe the color information of each frame. We quantize the hue to 30 levels and the saturation to 32 levels, therefore the dimension of HSH is 960. Similar to the IDT based feature, PCA is utilized to reduce the dimension of HSH feature. To encode the color information of a video, GMM is applied to construct a codebook. We compute one Fisher Vector over the current temporal scale, and set the cluster number of GMM and the temporal scale number to 512 and 2, respectively.

2.1.5 CNN Based Feature

In the MediaEval Violence Detection subtask, we also train a Convolutional Neural Network (CNN) to extract appearance features. CNN includes five convolution and pooling layers to extract appearance features and three layers of full connections for classification. CNN is well known for its powerful ability in feature extraction. However, CNN's generalization ability will be limited if there is not enough samples for training. Therefore, the images from the ImageNet dataset are used to pre-train the CNN. Frames from the Violence Detection subtask are used to fine-tune the first five layers and retrain the last three full connection layers. The architecture of CNN is the same as the CNN_M_2048 model [1].

2.2 Classification

As far as classification is concerned, the linear SVM is employed in this work. In addition, the One-Against-Rest approach is used for multi-label classification. In order to achieve the balance of training samples for each of the multiple classes, we weight positive and negative samples in an inverse manner. In our system, the standard linear LIBSVM is used with the penalty parameter C equal to 100. To combine different types of features, the late fusion strategy is utilized to linearly combine classifier scores computed for each feature.

3. RESULTS AND DISCUSSIONS

We submitted 5 runs with the results given in Table 1. The Violence Detection subtask of Run 1 used the IDT based feature, Dense SIFT feature, MFCC feature, HSH feature and CNN based feature. Run 2 and the Induced Affect Detection subtask of Run 1 used the IDT based feature, Dense SIFT feature, MFCC feature and HSH feature. Run 3 used the IDT based feature, Dense SIFT feature and MFCC feature. The Violence Detection subtask of Run 4 just used the CNN based feature. The Induced Affect Detection subtask of Run4 used the IDT based feature and Dense SIFT feature. The Violence Detection subtask of Run 5 used the IDT based feature, Dense SIFT feature, HSH feature and CNN based feature. The Induced Affect Detection subtask of Run 5 used the IDT based feature, Dense SIFT feature and HSH feature.

From the comparison, we can see that the motion cue is important for both subtasks. For the Violence Detection subtask, the comparison of Run 1 and Run 2 shows that the CNN based feature can help to improve the performance. For the Induced Affect Detection subtask, the comparison of Run 4 and Run 5 shows that the color information has a significant impact.

Table 1: Results of MIC-TJU.

	Arousal(%)	Valence(%)	Violence(%)
Run 1	55.93	41.95	28.48
Run 2	55.93	41.95	24.02
Run 3	55.61	40.81	21.76
Run 4	53.70	40.90	17.43
Run 5	55.32	40.92	26.47

We report average precision for Violence, global accuracy for Arousal and Valence [4].

4. REFERENCES

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC'14*, 2014.

[2] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab. `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`.

[3] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR'07*, 2007.

[4] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *MediaEval 2015 Workshop*, 2015.

[5] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV'13*, 2013.

[6] H. Wang, Y. Yi, and J. Wu. Human action recognition with trajectory based covariance descriptor in unconstrained videos. In *ACM MM'15*, 2015.