Fudan at MediaEval 2013: Violent Scenes Detection Using Motion Features and Part-Level Attributes

Qi Dai, Jian Tu, Ziqiang Shi, Yu-Gang Jiang, Xiangyang Xue
School of Computer Science, Fudan University, Shanghai
{daiqi,13210240062,09300240067,ygj,xyxue}@fudan.edu.cn

ABSTRACT

The Violent Scenes Detection Task of MediaEval provides a valuable platform for algorithm evaluation and performance comparison. This is a very challenging task as there exist many forms of violent scenes, which vary significantly in their visual and auditory clues. In this notebook paper, we describe our system used in MediaEval 2013, which focuses on the use of motion-based features and part-level semantic attributes. One of the key components of the system is a set of trajectory-based motion features that have been observed effective in last year's evaluation. We also adopt a newly developed part-level attribute feature, which consists of detection scores of object and scene parts. Our results indicate that the trajectory-based motion features can still offer very competitive performance, and the attribute feature is also helpful under several situations. In addition, temporally smoothing detection scores can lead to a significant performance gain. We conclude that a successful violent scenes detection system should use truly multimodal features, ranging from motion-based to static visual descriptors, as well as audio and attribute features.

1. INTRODUCTION

Detecting violent scenes in movies is a very interesting challenge that is receiving increasing attention in the vision and multimedia communities. The aim is to develop a fully automatic system that can reliably detect violent segments. An overview of this year's task, including data, labels and evaluation metrics, can be found in [1]. In this paper, we briefly introduce our system and discuss evaluation results.

2. SYSTEM DESCRIPTION

An overview of our system is shown in Figure 1. We first extract a comprehensive set of features, and then use SVM classifiers for violent scenes detection.

2.1 Feature Extraction

Four kinds of audio-visual features are extracted, including a set of trajectory-based motion features, spatial-temporal interest points (STIP), attribute features and audio features.

Trajectory-based Features: We compute trajectory-based motion features according to our recent work in [2]. We first compute the dense trajectories and extract histograms of oriented gradients (HOG), histograms of optical flow (HOF) and motion boundary histograms (MBH) on the spatial-temporal volumes along the trajectories. These three features, as well as a trajectory shape descriptor, are quantized using separate visual codebooks to generate four bag-of-words descriptors (4096 dimensions each). In order to improve the computation efficiency, we adopt Random Forest for quantization. Moreover, we also compute our proposed motion representation TrajMF [2] based on the motion relationships between trajectory pairs, using trajectory features HOG, HOF and MBH, respectively. As the dimension of the original TrajMF is very high, we employ expectation-maximization algorithm on principal component analysis (EM-PCA) for dimensionality reduction, leading to a 1500 dimensional representation for each feature. We have seven trajectory-based features in total, including four baseline bag-of-words and three dimension reduced TrajMF features. Readers are referred to [2] for more details.

STIP: Another popular motion feature is STIP [3], which is often used for video classification and action recognition. The STIP algorithm searches for interest points which have a dramatic change in both spatial and temporal dimensions, and computes HOG-HOF joint descriptors around the found points. Again, we adopt the bag-of-words framework, which converts the point descriptors to a histogram-style vector, using a 4000 dimensional codebook. The soft weighting strategy is used in the quantization process.

Part-Level Attributes: A new feature used in this year's system is part-level attributes, computed based on our recent work in [4]. The part filters are learned using the deformable part-based models [5], originally designed for object detection. Training data are collected from the ImageNet and the MIT scene datasets. The filters are then applied to videos frames[1], and the max response values at different scales are concatenated to form attribute descriptor for each frame, where each dimension has a certain semantic meaning, i.e., the likelihood of containing an object/scene part in the frame. Finally we generate an video-level attribute feature by applying the max-pooling strategy over all the extracted frames. An approximation method based on sparse coding to speed up the computation of filter maps is also adopted. For more details, please refer to [4]. Note that the sparse coding-based method is a new work and was not introduced in [4], which however does not prevent the understanding of the overall feature computation flow.

Mel-Frequency Cepstral Coefficients (MFCC): Many violent scenes have clear auditory clues. We extract the

[1]We sample one frame every two seconds.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain.

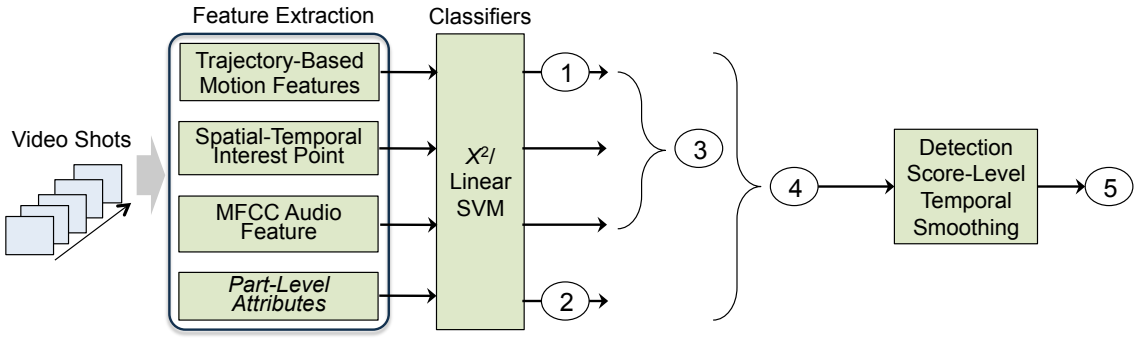

Figure 1: The framework of our system. Circled numbers indicate the 5 submitted runs.

well-known MFCC as the audio feature. The MFCC is extracted in every 32ms time-window with 50% overlap, using 20 cepstral coefficients. The bag-of-words framework is also employed here to quantize the MFCCs, using a codebook of 4000 words.

2.2 Temporal Score Smoothing

A violent scene segment may contain several continuous shots. While some shots can be identified as violent scenes easily, some of them may be hardly detected due to many issues. For instance, if only a small fraction of a shot contains violent scenes, the features of the violent parts may be dominated by that of other non-violent scenes. In this case, temporal information is important to mitigate the problem. In other words, if a shot contains a violent scene, its neighboring shots are relatively more likely to contain violence.

We exploit the temporal information using a very simple but efficient score smoothing method, where the smoothed prediction score of a shot is the average value of the violence scores over a three-shot window.

2.3 Classification

As mentioned earlier, we use SVM for classification. The widely used χ^2 kernel is adopted on the histogram-like features (the four baseline trajectory-based features and the STIP feature), while the linear kernel is used for the dimension reduced TrajMF features and the attribute feature.

2.4 Submitted Runs

As shown in Figure 1, we submitted 5 runs based on different combinations of the features. Run 1 and Run 2 are based on trajectory-based motion features and the part-level attributes respectively. Run 3 is the combination of Run 1 and the MFCC and STIP features. Run 4 further includes the attribute feature. Finally, Run 5 is generated by temporally smoothing the scores from Run 4. In all the submitted runs, kernel-level average fusion is used for combining the motion-based visual features, while score-level average late fusion is used in other cases.

3. RESULTS AND DISCUSSION

Figure 2 shows the performance of our submitted runs. As discussed in [1], there are two kinds of subtasks focusing on objective and subjective definitions of violence separately. We run our system on both kinds of labels. Overall, the results are quite promising. Results based on the subjective definition are generally better than those on the objective definitions, indicating that the subjectively determined violence may be more consistent in visual and auditory appearances. Attribute features are not as good as the motion-based ones, but the combination of them shows clear improvements on the subjective definition (from 0.571 to 0.587). We expect that the contribution of the attribute features can be largely enhanced if the part filters could be trained on similar data containing some violent scenes (in contrast to the current training data of Internet images with contents mostly unrelated to violence). Combining the trajectory-based motion features with MFCC and STIP leads to a big improvement. These results clearly show that using multimodal features is important in violent scenes detection. Finally, similar to our observations from last year's evaluation, the score smoothing is able to significantly improve the results.

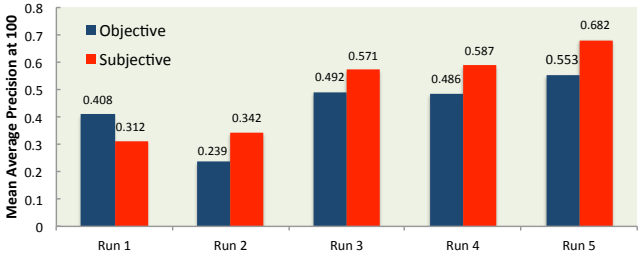

Figure 2: Performances of our 5 runs on both objective definitions (blue) and this year's newly added subjective definitions (red), measured by mean average precision over top 100 detected shots.

Acknowledgements

This work was supported in part by a National 973 Program (#2010CB327900), two grants from NSF China (#61201387 and #61228205), a grant from STCSM (#12XD1400900), and a New Teachers' Fund for Doctor Stations, Ministry of Education, China (#20120071120026).

4. REFERENCES

[1] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[2] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*, 2012.

[3] I. Laptev. On space-time interest points. *IJCV*, 64:107–123, 2005.

[4] Y. Zheng, Y.-G. Jiang, and X. Xue. Learning hybrid part filters for scene recognition. In *ECCV*, 2012.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. In *TPAMI*, 32(9):1627-1645, 2010.