The Shanghai-Hongkong Team at MediaEval2012: Violent Scene Detection Using Trajectory-based Features

Yu-Gang Jiang[§], Qi Dai[§], Chun Chet Tan[‡], Xiangyang Xue[§], Chong-Wah Ngo[‡]
[§]School of Computer Science, Fudan University, Shanghai
[‡]Department of Computer Science, City University of Hong Kong, Hong Kong
{ygj,daiqi,xyxue}@fudan.edu.cn
cctan2@student.cityu.edu.hk, cscwngo@cityu.edu.hk

ABSTRACT

The Violent Scene Detection task offers a very practical challenge in detecting complex and diverse violent video clips in movies. In this working note paper, we will briefly describe our system and discuss the results, which achieved top performance in mAP@20[1] and runner-up in mAP@100, among all 35 submissions worldwide.

The central component of our system is a set of features derived from the appearance and motion of local patch trajectories [2]. We use these features and SVM classifier as the baseline approach and add in a few other components to further improve the performance. Our findings indicate that the trajectory-based visual features already offer very competitive results. Other audio-visual features like Spatial-Temporal Interest Points and MFCC do not significantly enhance the performance. In addition, smoothing detection scores of nearby shots leads to significant improvement. We conclude that—while audio feature may help marginally—good visual features are still the key factor in violent scene detection, and temporal information is very useful.

Keywords

Violent scene detection, movie, trajectory-based feature, multi-modality, temporal smoothing.

1. INTRODUCTION

Automatically detecting violent scenes in videos has great potential in several applications, such as movie selection or recommendation for children. Figure 1 shows two examples of violent scenes in Hollywood movies. The annual MediaEval evaluation introduced this problem in 2011. An overview of this year's task can be found in [1].

In MediaEval 2012, we explored several interesting issues with a particular focus on novel features. We briefly describe each of the system components in the following.

2. SYSTEM DESCRIPTION

Figure 2 gives an overview of our system framework. We extract a diverse set of audio-visual features and use χ^2 kernel SVM classifier for violent scene detection.

[1]Mean average precision over top 20 detected shots.

Copyright is held by the authors/owners.
MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy.

Figure 1: Example frames of violent scenes in movies.

2.1 Feature Extraction

Trajectory-based Features: The trajectory-based features are computed based on our recent work [2]. Specifically, dense trajectories are first extracted using the method of [5], and each trajectory is described by three features, namely HOG, HOF and MBH. Based on the trajectories, we further generate motion representations called TrajMF by exploring relative locations and motions between trajectory pairs. In total we have seven features, including a trajectory shape feature, three bag-of-visual-words (BoW) features based on the three trajectory features respectively, and three TrajMF features [2]. These features serve as a strong baseline in our system. Readers are referred to [2] for more details.

SIFT: Two sparse keypoint detectors, Difference of Gaussian and Hessian Affine, are adopted to locate local invariant image patches from video frames. Each patch is described by a 128-d SIFT descriptor [4]. Since keypoint detection on every frame is computationally expensive and nearby frames are similar and redundant, we sample two frames per second. This feature is represented using the popular BoW framework, using two 500-d codebooks (generated from the two kinds of local patches separately).

Spatial-Temporal Interest Points (STIP): STIP captures a space-time volume in which video pixel values have large variations in both space and time. Laptev's algorithm is adopted [3]. This feature is also converted to BoW histograms, using a vocabulary of 4000 codewords.

MFCC: The well-known MFCCs are the only audio feature in our framework, which are computed for every 32ms time-window with 50% overlap. Again, the BoW framework is used to convert a set of MFCCs from each video shot into fixed dimensional vectors, using a vocabulary of 4000 audio codewords.

Concept-based Feature: Different from the low-level features described earlier, concept-based feature contains *mid-level* indicators where each dimension is the prediction output of a semantic concept. Ten concepts are provided in MediaEval2012, covering violence-related topics such as "presence of blood", "fights", "presence of fire", "gunshots", etc. We use the above low-level features to train SVM detec-

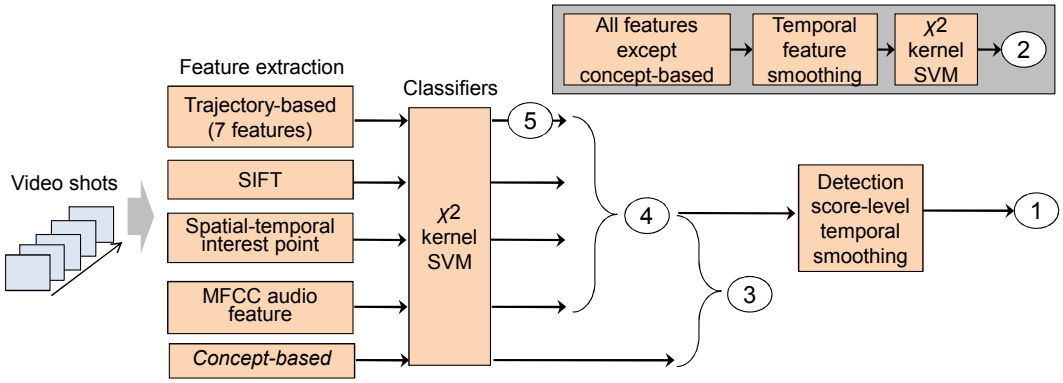

Figure 2: The framework of our violent scene detection system. Circled numbers indicate the 5 submitted runs.

tors for each of the concepts[2], and generate a concept-based representation of 10 dimensions for each video shot.

2.2 Temporal Smoothing

It is well-known that temporal structure is useful for video content analysis. There exist complex methods in the modeling of temporal information, such as the use of graphical models. In this task, we opt for a simple but very efficient temporal smoothing method, which takes clues from the shots before and after a target shot into account. Two smoothing choices are adopted.

Feature Smoothing: This uses the averaged feature over a three-shot window to represent the shot in the middle of the window. Classification/Detection is performed on the smoothed features.

Score Smoothing: Different from feature smoothing, score smoothing uses features from each single shot for classification, and smooth (average) prediction scores over three-shot windows.

2.3 Submitted Runs

As indicated in Figure 2, we submitted five runs based on different feature combinations and/or smoothing choices. Our baseline run 5 uses the seven trajectory-based features, and run 4 includes three additional features (SIFT, STIP and MFCC). Run 3 further combines the concept-based feature. Run 2 and run 1 are generated by the two temporal smoothing methods respectively, using the same feature set to run 4. In all the submitted runs, kernel-level early fusion (mean of the individual-feature kernels) is used to combine multiple features.

3. RESULTS AND DISCUSSION

Figure 3 shows the performance of all the 35 official submissions, where our run 1 produces the highest mAP@20 accuracy (0.736). Our run 5, which uses only the seven trajectory-based features, already shows very competitive results (mAP@20=0.656, mAP@100=0.539). This is very appealing since the features were initially designed for detecting simple human actions [2]. However, the three additional features used in run 4 are not helpful (mAP@20=0.666, mAP@100=0.508). This may be due to an implementation issue in the SIFT representation, which is still under investigation. In addition, the concept-based scores (run 3) do not contribute as well (mAP@20=0.650, mAP@100=0.502). One possible reason is that the concept detectors are trained on the same set of training videos, which would generate detection scores in different scales between the training and test sets, because classifiers like SVM always tend to overfit training data. The resulted 10-d features of different scales largely limit the contribution of this concept-based representation. Nevertheless, we believe that—if there are separate and sufficient training data—using mid-level concept detectors is a promising direction to further improve violent scene detection accuracy.

[2]For the three audio concepts, only MFCC features are adopted.

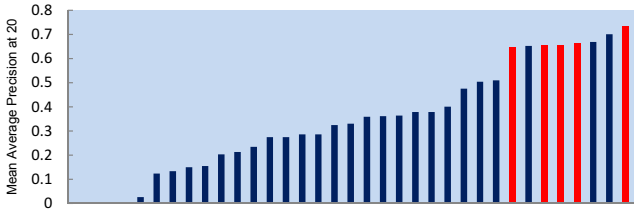

Figure 3: Performance of our 5 submitted runs (red) and all the other 30 official submissions (blue), measured by mean average precision (mAP) over top 20 detected shots.

Comparing the two temporal smoothing choices, score smoothing is significantly better (run 1: mAP@20=0.736, mAP@100=0.624; run 2: mAP@20=0.654, mAP@100=0.561), indicating that *blurring* features across shots is not a good option.

Acknowledgements

This work was supported in part by two grants from the National Natural Science Foundation of China (#61201387 and #61228205), and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709).

4. REFERENCES

[1] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection. In *MediaEval 2012 Workshop*, Pisa, Italy, 2012.

[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] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[5] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.