NII-UIT at MediaEval 2014 Violent Scenes Detection Affect Task

Vu Lam
University of Science
227 Nguyen Van Cu, Dist.5
Ho Chi Minh, Vietnam
lqvu@fit.hcmus.edu.vn

Duy-Dinh Le
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan 101-8430
ledduy@nii.ac.jp

Sang Phan
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan 101-8430
plsang@nii.ac.jp

Shin'ichi Satoh
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan 101-8430
satoh@nii.ac.jp

Duc Anh Duong
University of Information Technology
KM20 Ha Noi highway, Linh Trung Ward,Thu Duc District
Ho Chi Minh, Vietnam
ducda@uit.edu.vn

ABSTRACT

Violent scene detection (VSD) is a challenging problem because of the heterogeneous content, large variations in video quality, and semantic meaning of the concepts. The Violent Scenes Detection Task of MediaEval [1] provides a common dataset and evaluation protocol thus enables a fair comparison of methods. In this paper, we describe our VSD system used in MediaEval 2014 and briefly discuss the performance results obtained in main subjective tasks. In this year, we focus on improving the trajectory-based motion features that have been proven effective in previous year's evaluation. Besides that, we also adopt SIFT-based and audio features as in last year's system. We combined these features using late fusion. Our results show that the trajectory-based motion features still have very competitive performance and the combination with still image features and audio features can improve overall performance.

1. INTRODUCTION

We consider the Violent Scenes Detection (VSD) task [1] as a concept detection task. For evaluation, we use our NII-KAORI-SECODE framework, which has been achieved good performances on other benchmarks such as ImageCLEF and PASCALVOC. Firstly, videos are divided into equal segments with 5-second length. In each segment, keyframes are extracted by sampling 5 keyframes per second. For still image features, local descriptors are extracted and encoded for all keyframes in each segment and then segment-based features are formed from their keyframe-based features by applying average or max pooling. Motion feature and audio feature are extracted directly from the whole segment. For all features, we use the popular SVM algorithm for learning. Finally, the probability output scores of the learned classifier are used for ranking retrieved segments.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

2. FEATURE EXTRACTION

We use features from different modalities to test if they are complementary for violent scenes detection. Currently, we have developed our VSD system to incorporate still image feature, motion feature, and audio feature.

2.1 Still Image Features

In this year, we use only SIFT-based features for VSD because they could capture different characteristics of images. We use popular SIFT-based features with both Hessian Laplace interest points and dense sampling at multiple scales. Besides the standard SIFT descriptor, we also use Opponent-SIFT and Color-SIFT [2]. We employ the bag-of-words model with a codebook size of 1000 and the soft-assignment technique to generate a fixed-dimension feature representation for each keyframe. Beside encoding the whole image, we also divide it into grids of 3x1 and 2x2 to encode spatial information. Finally, in order to generate a single representation for each segment, we use two pooling strategies: average pooling and max pooling.

2.2 Motion Feature

We use the Improved Trajectories [3] to extract dense trajectories. A combination of Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) is used to describe each trajectory. We encode HOGHOF and MBH features separately using the Fisher Vector encoding. The codebook size is 256, trained using a Gaussian Mixture Model (GMM). The feature representation of each descriptor after applying PCA has 65,536 dimensions. Finally, these two features are concatenated to form the final feature vector with 131,072 dimensions.

2.3 Audio Feature

We use the popular Mel-frequency Cepstral Coefficients (MFCC) for extracting audio features. We choose a length of 25ms for audio segments and a step size of 10ms. The 13-dimensional MFCC vectors along with each first and second

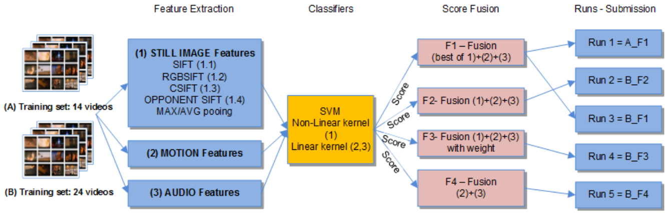

Figure 1: Overview of our system and the 5 submitted runs.

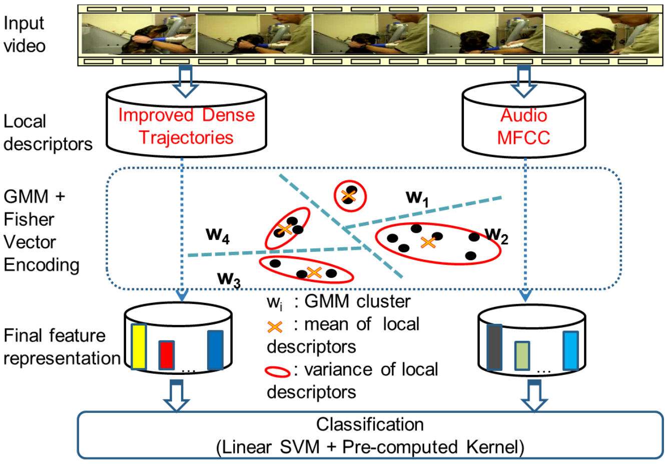

Figure 2: Our framework for extracting and encoding motion and audio feature.

derivatives are used for representing each audio segment. Raw MFCC features are also encoded using Fisher vector encoding. We use a GMM to train the codebook with 256 clusters. For audio features, we do not use PCA. The final feature descriptor has 19,968 dimensions. Our motion and audio framework are shown in Fig 2.

3. CLASSIFICATION

LibSVM [4] is used for training and testing at segment level. To generate training data, segments of which at least 80% are marked as violent according to the ground truth. Extracted features are scaled to $[0, 1]$ using the SVM-scale tool of LibSVM. The remaining segments are considered as negative. For still image features, we use a chi-square kernel to calculate the distance matrix. For audio and motion features, which are encoded using Fisher vector, a linear kernel is used. The optimal gamma and cost parameters for learning SVM classifiers are found by conducting a grid search with 5-fold cross validation on the training dataset.

4. SUBMITTED RUNS

We select two training sets: (A) uses 14 videos, (B) uses 24 videos. We use the VSD 2013 test dataset (7 videos) as validation set. We employ a simple late fusion strategy on the above features, using equal weights and learnt weights. We submitted five runs in total (Fig 1): (R1) using training set A, we first select the best still image feature and fuse it with motion and audio features; (R2) using training set B, we fuse all still image features with motion and audio using equal weight; (R3) same as R1 but using training set B; (R4) using training set B, we fuse all still image features with motion and audio using learnt fusion weights from validation set; (R5) using training set B, we fuse motion and audio features with equal weights.

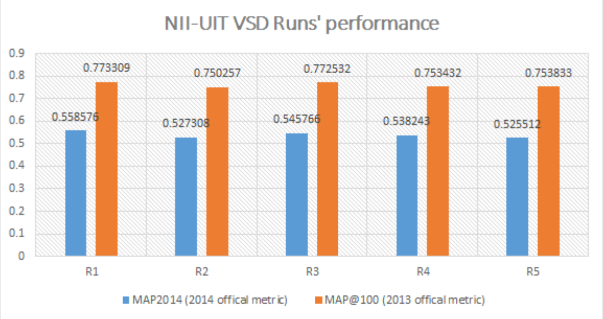

Figure 3: Results for the main task with MAP2014 and MAP@100(2013) metrics

5. RESULTS AND DISCUSSIONS

The detailed performance for each submitted run is shown in Figure 3. Our best run is the fusion run of best single still image features (RGBSIFT), motion and audio features (R1). There is not a big gap among submitted runs. We see that, the performance of motion features with Fisher vector encoding is alway good and significantly better than others. In all submitted runs, we used motion features as a base to fuse with others. Audio and still image features did not achieve good performance, but they can be complementary to motion features. Another interesting observation is that runs trained on fewer videos (training set A - 14 videos) have better performance than the runs in which set (24 videos) was used. This indicates that the second training set might contain ambiguous violent scene's annotations, which harms the detection performance.

6. ACKNOWLEDGEMENTS

This research is partially funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-26-01.

7. REFERENCES

[1] M. Sjöberg, B. Ionescu, Y. Jiang, V. Quang, M. Schedl, and C. Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In MediaEval 2014 Workshop, Barcelona, Spain, October 16-17 2014.

[2] K. Van de Sande, T. Gevers, C. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.32, no.9, pp.1582,1596, Sept. 2010

[3] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV '13). IEEE Computer Society, Washington, DC, USA, 3551-3558.

[4] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.