NII-UIT at MediaEval 2013 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

We present a comprehensive evaluation of shot-based visual and audio features for MediaEval 2013 - Violent Scenes Detection Affect Task. To obtain visual features, we use global features, local SIFT features and motion features. For audio features, the popular MFCC is employed. Besides that, we also evaluate the performance of mid-level features which is constructed using visual concepts. We combined these features using late fusion. The results obtained by our runs are presented.

Keywords

semantic concept detection, global feature, local feature, motion feature, audio feature, mid-level feature, late fusion

1. INTRODUCTION

We have developed NII-KAORI-SECODE, a general framework for semantic concept detection, and used it to participate in several benchmarks such as IMAGECLEF, MEDIAEVAL, PASCAL-VOC, IMAGE-NET and TRECVID. In this year, we evaluate performance for concept detection-like task using shot-based feature representations only. Our previous works show that using the shot-based features not only reduce the computational cost but also improve the performance.

We consider the Violent Scenes Detection (VSD) Task [1] as a concept detection task and use the NII-KAORI-SECODE framework for evaluation. Firstly, keyframes are extracted by sampling 5 keyframes/second. Raw features are extracted for all keyframes in each shot and then shot-based features are formed from its keyfame-based feature by applying average or max pooling. Motion feature and audio feature are extracted directly from the whole shot. For mid-level feature, at first we build attribute classifiers for 7 visual attributes: fights, blood, gore, fire, car chase, cold arms, firearms. After that, we concatenate output scores of each attribute classifier to form the mid-level feature representation. 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 shots.

We use the same framework for evaluating both objective and subjective tasks (just different annotations). Our results show that the combined runs using all visual, audio and mid-level features achieved the best performance.

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

2. LOW LEVEL FEATURE

We use feature 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

We use both global and local features for VSD because they capture different characteristics of images. For global feature, we use Color Histogram (CH), Color Moment (CM), Edge Oriented Histogram (EOH), and Local Binary Pattern (LBP). For local feature, we use popular SIFT with both Hessian Laplace interest points and dense sampling at multiple scales. For dense sampling, besides the standard SIFT descriptor, we also use Opponent-SIFT and C-SIFT. For interest point detector, we only use normal SIFT descriptor. We also employed 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 divided it into grids of 3x1 and 2x2 to encode spatial information. Finally, in order to generate a single representation for each shot, we employed two pooling strategies: average pooling and max pooling.

2.2 Motion Feature

Trajectories are obtained by tracking the densely sampled points in the optical flow fields. We use Motion Boundary Histogram (MBH) to describe each trajectory. This feature descriptor is know to perform well for handling camera motion. For motion feature we use Fisher vector encoding after reducing feature dimension using PCA. The codebook size is 256, trained using a Gaussian Mixture Model (GMM). The final feature dimension is 65,536.

2.3 Audio Feature

We use the popular MFCC for extracting audio feature. We choose a length of 25ms for audio segments and a step size of 10ms. The 13d MFCCs along with each first and second derivatives are used for representing each audio segment. Raw MFCC features are also encoded using Fisher vector. We use GMM to build the codebook with 256 clusters. We also apply PCA to reduce feature dimension, resulting feature descriptors of 12,288 dimensions.

3. MID-LEVEL FEATURE

Beside low-level features, we also investigate how to use related violent information as mid-level feature to detect violent scenes. We use only seven violent concepts to create attributes: fire, firearms, cold arms, car chase, gore, blood, and fight. We use low-level image feature to train the attribute classifiers on the VSD development set of 2011. For each image, we apply these attribute classifiers to get score values corresponding to each attribute. After that, we concatenate all these values to form the mid-level representation of each image. We then train our mid-level classifier on the VSD development and test set of 2012. Finally, this classifier is used for testing on this year's set. The detailed work flow is shown in Figure 1.

4. CLASSIFICATION

LibSVM is used for training and testing at shot level (based on shot boundaries provided by the organizers). To generate training data, shots which fall into positive segments more than 80% will be considered as positive shots. The remaining shots are considered as negative. Extracted features are scaled to $[0, 1]$ using the svm-scale tool of LibSVM. For still image features, we use a chi-square kernel to calculate the distance matrix. For audio and motion feature, 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.

5. SUBMITTED RUNS

We employ a simple late fusion strategy on the aforementioned low-level and mid-level features, giving equal weights to the different factors. We submitted five runs in total: (R5) Fusion of all 4 global features and 5 local features; (R4) Fusion of motion feature (dense trajectories + MBH) and audio feature (MFCC); (R3) The run using mid-level feature; (R2) Fusion of R4 and R5; and (R1) Fusion of R3, R4 and R5.

6. RESULTS AND DISCUSSIONS

The detailed performance for each submitted run is shown in Figure 2. We report the performance of both objective and subjective tasks. For each task, we report two evaluation metrics: overall MAP and MAP100, which is the MAP at top 100 return shots. Our best run is the fusion run of all global, local, motion and audio feature (R1). This observation confirm the benefit of combining multiple features for violent scenes detection. Among all submitted runs, the run using mid-level (R3) performs the worst. However, it can be complementary for combining with other low-level features (R1). The combined run using motion feature and audio (R4) feature did not achieve good results as expected. In fact, its performance is lower than the combined run of still image features (R5). This can be due to minor motion in each shots and/or noise in audio signals.

Our future study includes investigating the contribution of motion features and audio features. The result of mid-level features is also promising. Currently, we only use 7 visual concepts for constructing mid-level features. In the future, we will incorporate audio concepts using audio feature.

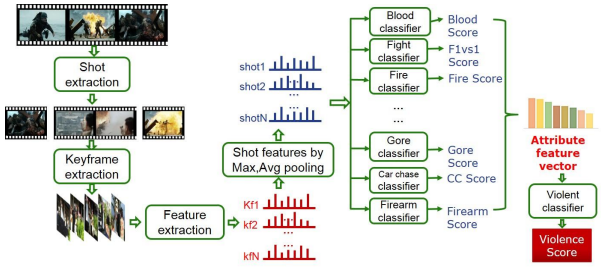

Figure 1: Mid-level feature construction

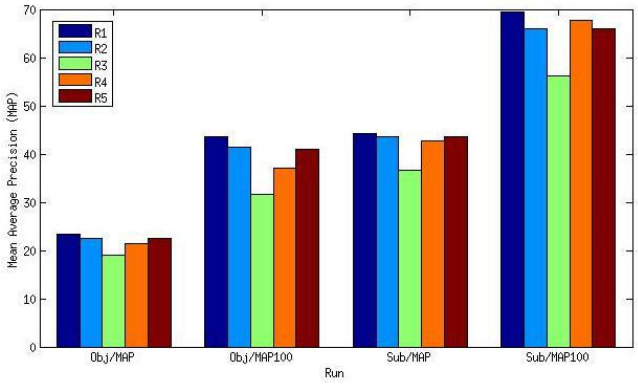

Figure 2: Results of our submitted runs

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] Demarty C.H, Penet C., Schedl M., Ionescu B., Lam Q. V. and Jiang Y. G. *The MediaEval 2013 Affect Task: Violent Scenes Detection*, MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain.