NII, Japan at MediaEval 2012 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 Le
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 performance of shot-based visual feature representations for MediaEval 2012 - Violent Scenes Detection Affect Task. In spite of using keyframe-based as last year, we try to apply shot-based features using the global features (color moments, color histogram, edge orientation histogram, and local binary patterns) for violent scenes detection. Besides that, we also evaluate the performance of late fusion with visual attributes (blood, fights, gore, car chase, fire, coldarm, firearm). The results obtained by our runs are presented.

Keywords

semantic concept detection, shot-based feature, global features, late fusion, attributes

1. INTRODUCTION

We have developed NII-KAORI-SECODE, a general framework for semantic concept detection, and used it to participate several benchmarks such as IMAGECLEF, MEDIAEVAL, PASCAL-VOC, IMAGE-NET and TRECVID. In this year, we try to evaluate performance of shot-based visual feature representations for concept detection-like task.

In this framework, first the keyframes will be extracted from movies based on shot boundary (we selected 5 keyframes per shot), features are extracted from keyframes, keyframe features will be formed shot features (based on shot provided by Mediaeval) by average, max, min pooling, then concept detectors (for violent concepts and 7 visual attributes: fights, blood, gore, fire, carchase, coldarms, firearms) using these keyfame-features and shot-features are learned by using SVM with RBF kernel.

The probability output scores of the learned concept detectors are used for ranking both keyframes and shots. We consider the Violent Scenes Detection Task [1] as a kind of concept detection task and use NII-KAORI-SECODE framework for evaluation of performance of keyframe-based and shot-based features.

We also evaluate the performance of using 7 visual attributes and try to investigate how to use these attributes to enhance the performance of violent detectors. The results using shot-based features are more effective than using keyframe-based features and combination of visual attributes can not help to improve the overall performance.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2. FEATURE EXTRACTION

We evaluate both keyframe-based features and shot-based features by using global features. The global features include color moments, color histogram, edge orientation histogram, and local binary patterns.

2.1 Feature Configuration

2.1.1 Granularity

Since global features do not capture spatial information, to overcome this problem, a grid $n \times m$ is usually used to divide the input image into non overlapping sub-regions. The features extracted from these regions are concatenated to form the feature vector for the image.

2.1.2 Color space

Local binary patterns and edge orientation histogram are extracted from gray scale image. For color moments and color histogram, color spaces including HSV, RGB, Luv, and YCrCb are used.

2.1.3 Quantization

For color histogram, we only use 8-bin histogram for each channel. For edge orientation histogram, we quantize orientations into histograms of 12+1 bins, 18+1 bins, 36+1 bins, and 72+1 bins. For local binary patterns, we quantize binary patterns into histograms of 10, 30, and 59 bins.

Table 1: Performance of NII's runs (sorted by MAP)

Participant	AED COST	MAP	MAPAT100
Run3-nsc.cCV-GRAY.g4.q59.g-lbp.shotmax	1.286798	0.15265	0.30829
Run2-GlobalFeatures.shotmax	1	0.14882	0.27638
Run1-GlobalFeatures.shotmax-violentscenes-fights-blood-gore	1	0.14867	0.245
Run4-nsc.cCV-GRAY.g4.q30.g-lbp.shotmax	1.860452	0.1235	0.22441
Run5-nsc.cCV-Luv.g5.q8.g-ch.shotmax	2.113624	0.10287	0.148107

Each combination of feature type, granularity, quantization, and color space forms one feature configuration.

2.2 Keyframe-based and Shot-based features

We extracted the features for each keyframe (using feature configurations in previous section). For each shot, 5 keyframes were extracted then we have 5 feature vectors, we formed the feature vector for each shot from these vectors by MAX, MIN, AVERAGE pooling and CONCATENATION 5 keyframe's vectors

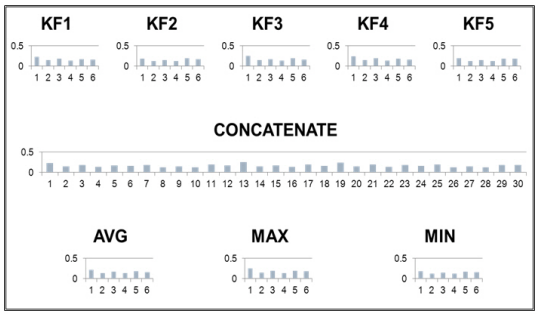

Figure 1: Feature vectors for each SHOT

3. CLASSIFIER LEARNING

LibSVM is used to train SVM classifiers. The extracted features are scaled to $[0, 1]$ using the svm-scale tool of LibSVM. The RBF kernel (with chi-square distance) is used as similarity measure. The optimal (C, g) parameters for learning SVM classifiers are found by conducting a grid search with 5-fold cross validation on the original dataset.

4. EXPERIMENT

For each shot, 5 keyframes are extracted (based on shot boundary provided by organizers). Both keyframe feature and shot feature are used for training and testing. To generate training data, keyframes falling into positive segments will be considered as positive keyframes, shots which fall into positive segments more than 50 percentages will be considered as positive shots. The other keyframes and shots are considered as negative. We trained the classifiers for the violent concept and 7 visual attributes. We apply the trained classifier to the keyframes and shots of the test set.

The output scores of keyframes and shots are considered as scores of shots and used for ranking. We use the threshold $\theta = 0.02$ for binary decision. Besides that, we also have the classifiers for 7 visual attributes. We apply these classifies for the test set and got the score the shot in test set. We apply the fusion between the results of violent classifier and the results of 7 visual attributes classifiers.

We submitted 5 runs and the details of performances are shown in Table 1. AED cost is the cost defined by the task organizers, MAP is mean average precision, MAPAT100 is MAP of top 100 shots.

The results (in Figure 2) show that using shot-based features are more effective than using keyframe-based features, shot features by MAX pooling gives the best result and combination of violent concept and visual attributes cannot help to improve the overall performance

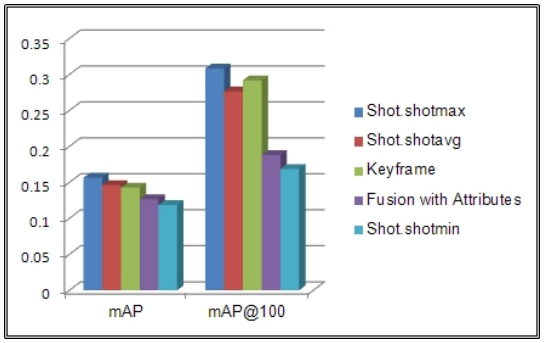

Figure 2: mAP comparision among keyframe-based, shot-based (min,max,avg), fusion with attributes of nsc.cCV-GRAY.g4.q30.g-lbp

5. DISCUSSION

The definition of violence is so general that we cannot know how to use the visual attributes to define the violent concept. We also try to apply the new annotation way to reduce the noisy keyframes. However, the length of shots are very diverse (from 3 to more than 10.000 keyframes), many shots are very short, and might be easily classified as non-violent shots based on the definition.

Our methods only work with the violent films. For non-violent films, the peformance is worse than the violent-films. Future work is to study how to use visual attributes to represent violent scenes. Fusion of the violence detection results with other visual attributes results cannot improve the performance.

6. REFERENCES

[1] Demarty C.H, Penet C., Gravier G. and Soleymani M. *The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies*, MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy.