NII, Japan at MediaEval 2011 Violent Scenes Detection 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

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 Science
227 Nguyen Van Cu, Dist.5
Ho Chi Minh, Vietnam
daduc@hcmus.edu.vn

ABSTRACT

We present a comprehensive evaluation of performance of visual feature representations for MediaEval 2011 - Violent Scenes Detection Task. As for global features, color moments, color histogram, edge orientation histogram, and local binary patterns are used. As for local features, keypoint detectors such as Harris Laplace, Hessian Laplace, Harris Affine, Dense Sampling are used to extract keypoints and SIFT and COLOR SIFT are used as descriptor of the region around these keypoints. The results obtained by our runs are presented. The demo is available at:
http://satoh-lab.ex.nii.ac.jp/users/ledduy/Demo-MediaEval

Keywords

semantic concept detection, local features, global features, sift, color sift

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. The purpose is to evaluate performance of various visual feature representations for concept detection-like task. In this framework, first features are extracted from keyframes, then concept detectors using these features are learned by using SVM with χ^2RBF kernel. The probability output scores of the learned concept detectors are used for ranking. 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 global and local features. The results show that using global features are more effective than using local features and combination of global features and local features can help to improve the overall performance.

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

2. FEATURE EXTRACTION

We evaluate both global features and local features. The global features include color moments, color histogram, edge orientation histogram, and local binary patterns. The local feature is based on the BOW model in which the SIFT descriptor is extracted at interest points detected by Harris Hessian Laplace and multi-scale dense sampling detector.

3. FEATURE CONFIGURATION

3.1 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.

3.2 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.

3.3 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.

Each combination of feature type, granularity, quantization, and color space forms one feature configuration. The feature configurations evaluated in this study are described in Table 1.

Table 1: Feature configurations.

Feature Type	Granularity	Color Space	Quantization (#Bins)	Total Configs
Color moments (CM)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	3x3	20
Color histogram (CH)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	8x3	20
Local binary patterns (LBP)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	10, 30, 59	15
Edge orientation histogram (EOH)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	12, 18, 36, 72	20
Local features (harhes, harlap, heslap, haraff, hesaff, dense, phow csift, sift, oppsift, rgbsift, rgsift)	1x1, 2x2, 1x3, 3x1	GRAY	500 visual words	44

4. 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 χ^2RBF kernel 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 a subset of 3,000 samples stratified selected from the original dataset.

5. EXPERIMENT

For each shot, one keyframe is used for training and testing. The set of keyframes are provided by the organizer. To generate training data, shots falling into positive segments are considered as positive shots. The other shots are considered as negative shots. We apply the trained classifier to the keyframes of the test set. The output scores of keyframes are considered as scores of shots and used for ranking. We use the threshold $\theta = 0.02$ for binary decision.

We submitted 6 runs and the details of performances are shown in Table 2. AED cost is the cost defined by the task's organizer and MAP is mean average precision. The results show that using global features are more effective than using local features and combination of global features and local features can help to improve the overall performance.

Figure 1 shows ROC curves of the submitted runs. The details of other runs are available at http://satoh-lab.ex.nii.ac.jp/users/ledduy/Demo-MediaEval.

Table 2: Performance of NII's runs(sorted by MAP)

RunID	Description	AED Cost	MAP (%)
NII-run6	Global features	1.000	25.13
NII-run5	Fusion of local and global features	1.000	24.79
NII-run1	LocalFeature-VLFEAT (DSIFT/PHOW)	1.947	17.22
NII-run4	LocalFeature-All3 (VLFEAT+COLORSIFT+VGG)	1.947	14.06
NII-run3	LocalFeature-COLORSIFT (sift,oppsift, rgsift,csift,rgbsift)	1.975	13.12
NII-run2	LocalFeature-VGG(harhes, harlap,haraff,heslap,hesaff)	1.976	12.87

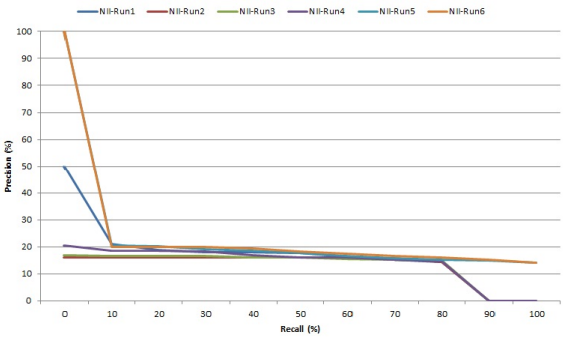

Figure 1: ROCCurve of the submitted runs.

6. DISCUSSION

- Using one keyframe per shot for training and testing needs to be reconsidered since in many cases, keyframes are not related to violent scenes. We tried using 10 keyframes per shot for training and testing. However, the performance is even worse than that of using only one keyframe. The reason could be multiple keyframes per shot in training data make large variations.
- Annotation and ground truth are based on segments while shots are used for experiments. Many shots are very short, e.g. less than 1 second and might be easily classified as non-violent shots based on the definition. Future work is to study how to use multiple keyframes per shot to represent violent scenes. Simple sampling does not work.

7. REFERENCES

[1] Demarty C.H, Penet C., Gravier G. and Soleymani M. *The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies*, MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy.