TUDCL at MediaEval 2013 Violent Scenes Detection: Training with Multi-modal Features by MKL

Shinichi Goto[1], Terumasa Aoki[1, 2]
1 Graduate School of Information
2 New Industry Creation Hachery Center
Tohoku University, Miyagi, Japan
{s-goto, aoki}@riec.tohoku.ac.jp

ABSTRACT

The purpose of this paper is to describe the work carried out for the Violent Scenes Detection task at MediaEval 2013 by team TUDCL. Our work is based on the combination of visual, temporal and audio features with machine learning at segment-level. Block-saliency-map based dense trajectory is proposed for visual and temporal features, and MFCC and delta-MFCC is used for audio features. For the classification, Multiple Kernel Learning is applied, which is effective if multi-modal features exist.

1. INTRODUCTION

The MediaEval 2013 Affect Task [1] is intended to detect violence scenes in movies. Although two different definitions of violent events are provided this year, our algorithm is developed only to solve the task for the objective definition, which is "physical violence or accident resulting in human injury or pain."

2. APPROACH

Rather than focusing on video shots from the beginning, our approach first handles fixed-length segments, each of which has 20 frames (0.8 seconds if FPS is 25). After segment-based scores are calculated from extracted feature vectors by machine learning, shot-based scores are generated.

For our runs only violent and non-violent ground truth are used, and neither a high-level concept nor external data is used.

2.1 Visual and Temporal Features

Both visual and temporal features based on dense trajectory [2] are calculated at every frame. Although the original dense trajectory algorithm is carried out by sampling frames densely except for homogeneous image areas, we additionally apply saliency maps proposed by Itti [3] to increase the precision, supposing that events concerned with violence are located in the areas people tend to pay attention to.

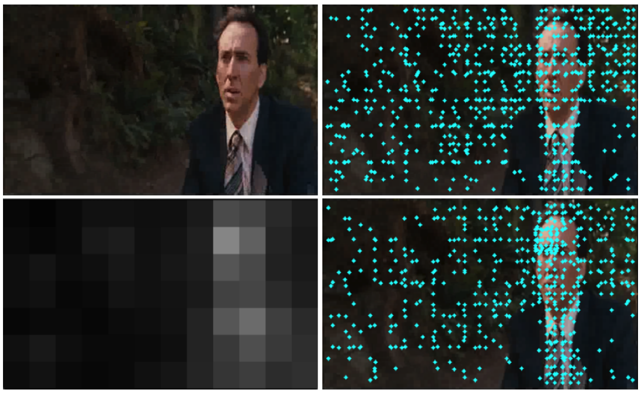

Figure 1: Example of dense sampling using saliency map: Original image (upper left), Normal dense sampling (upper right), Block saliency map (bottom left), Our dense sampling (bottom right).

In our algorithm, first a normal saliency map is generated, and then it is transformed to a block-based map by taking the average of salient values in a fixed block area so that dense sampling can be applied, changing its sampling step size and maximum spatial scale level according to the salient level. For instance, the most salient area in a image is densely sampled with the smallest step size, which guarantees the more salient a block is, the more points are obtained there. Figure 1 shows one example of our dense sampling and normal dense sampling. You notice that our algorithm is sampling more points in salient regions and less points in non-salient regions, but normal dense sampling, on the other hand, is taking points more uniformly on a whole frame. Note points in the homogeneous areas have already been deleted.

Trajectories, MBH, and additionally RGB histogram around trajectories are extracted for visual and temporal information, though in [2] HOG and HOF are also proposed. This is due to the fact that those features have poor contribution on our test runs.

All features are converted to Bag-of-Words form in each segment to get 200-d trajectory, 200-d MBH-x, 200-d MBH-y, and 400-d RGB histogram. In total, 1000-d feature vector is used as the visual and temporal feature for classification.

2.2 Audio Features

Major MFCC, delta-MFCC and audio energy is calculated every 20ms with 10ms overlap to create 200-d Bag-of-Audio-Words in each segment, which has 0.8 seconds.

2.3 Classifier Learning

Although a conventional way of tackling this classifying problem is to use Support Vector Machine (SVM), we apply Multiple Kernel Learning (MKL), which aims at finding

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

Table 1: Weights difference learned by MKL.

movie	Audio	Traj.	MBHx	MBHy	RGB
Armageddon	0.307	0.319	0.359	0.373	0.350
The Sixth Sense	0.450	0.180	0.407	0.440	0.171
Dead Poet Society	0.297	0.267	0.425	0.462	0.286

optimized weights when multiple SVM kernels are applied [4]. This suits well our case since multiple feature spaces exist. The whole kernel is composed of multiple kernels, and is computed according to the following equation:

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_k d_k \boldsymbol{K}_k(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad (1)$$

where $\boldsymbol{K}_k$ are base kernels, and d_k is a weight for each kernel. In our case, kernels for trajectory, x-direction MBH, y-direction MBH, RGB-histogram and audio features are prepared. For a kernel function, Histogram Intersection Kernel (HIK) is used since all of our features are histogram-based.

Although MKL can find optimal weights, we found these values are different depending on movies. Table 1 shows the difference between weights learned from three different movies. Therefore first classifiers for training movies are learned separately to give binary classification for each segment, and finally they are integrated in the following way.

2.4 Integration

The first step here is to calculate a pre-final violence score for each segment. To do so, for each segment in test movies, we simply calculate the number of classifiers which classify that segment as violent. Therefore for each test movie, a score s_i for the ith segment is:

$$s_i = \sum_{m=0}^{M-1} c_i(m), \quad c_i(n) = \{0, 1\} \quad (n = 0, 1, \ldots, M-1) \quad (2)$$

where $c_i(n)$ is a result of binary classification by the nth classifier with 0 for non-violence, 1 for violence. Note M is the total number of classifiers, which is equal to the number of training movies.

Finally a moving average is calculated as smoothing method for each test movie in order to decide final scores s'_t for all segments following:

$$s'_i = \frac{s_i + \sum_{n=1}^{N} \alpha^n \cdot (s_{i-n} + s_{i+n})}{2N+1} \quad (0 < \alpha < 1) \quad (3)$$

where α is a smoothing coefficient, N is a neighbor range around a segment. We used 0.5 for α and 2 for N.

The reason why this integration process is needed is to take the continuity of segments into account. Besides, since our classifier is learning each training movie separately, the violence concepts which a training movie does not have can be easily missed. Scores for shots are calculated by converting segment-based scores after calculating score per frame. If this score is higher than a threshold, that segment or shot is classified as violent. We choose 0.1 for a segment threshold, and 0.03 and 0.06 for shot thresholds.

3. RESULTS AND DISCUSSION

Shot-based results of our runs are shown in Table 2, and segment-based results are shown in Table 3. The difference between mkl-shot-hik-1 and mkl-shot-hik-2 is the value of the scoring threshold (0.03 for the former, 0.06 for the latter), and therefore it doesn't affect MAP@100. In addition to our main runs, results by normal SVM with RBF kernel are displayed for comparison, although there is no MAP@100 score since only binary classification results are decided and no score is calculated for SVM.

Table 2: Results of shot-level runs (Note all of them are AED metrics).

Run	MAP@100	Prec.	Rec.	F-sc.
mkl-shot-hik-1	0.470	0.222	0.726	0.340
mkl-shot-hik-2	0.470	0.284	0.609	0.387
svm-shot-rbf	-	0.0976	0.738	0.172

Table 3: Results of segment-level runs.

Run	MAP@100	Prec.	Rec.	F-sc.
mkl-seg-hik	0.343	0.214	0.309	0.253
svm-seg-rbf	-	0.0473	0.466	0.0859

Our results show the approach of Multiple Kernel Learning with HIK kernel is effective for violent scenes detection, though its F-score is still not high enough. We investigated this and came to the presumption that segments which have frequent camera motions, multiple people and loud sound tend to be mis-classified as violent.

On the other hand, common missed violent segments are violent scenes without sound, such as a scene in which a man is wringing on an another man's neck. It is reasonable to suppose that segments in which multi-modality cannot be exploited are likely to get missed.

Although MBH, which is proposed as robust to camera motions, is extracted, trajectories themselves easily get affected by camera motions, making them unreliable. Therefore some action against this problem is imperative.

It also should be added as classifiers have learned each training movie separately, feature vectors might not be enough compared to the case in which classifiers learn all movies simultaneously. Since not enough comparison with other methods have been done, we will continue our investigation.

4. REFERENCES

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

[2] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[3] Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 20 Issue 11*, pages 1254–1259, IEEE Computer Society Washington, DC, USA, November 1988.

[4] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. In *Journal of Machine Learning Research 5*, pages 27–72, 2004.