The Vireo Team at MediaEval 2013: Violent Scenes Detection by Mid-level Concepts Learnt from Youtube

Chun Chet Tan
Department of Computer Science
City University of Hong Kong, Hong Kong
cctan2-c@my.city.edu.hk

Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong, Hong Kong
cscwngo@cityu.edu.hk

ABSTRACT

The Violent Scenes Detection task continues to pose challenge in detecting violent scenes in Hollywood movies. In this working notes paper, we present the framework of our system and briefly discuss the performance results obtained in both objective and subjective subtasks. Besides using the low-level features for training the SVM classifiers for violent scenes detection, we show the feasibility in using the concept detectors to infer the occurrence of violent scenes. External Youtube data is exploited in our implementation to provide more diverse definition to violent scene concepts. Furthermore, we explore the feasibility of using Conditional Random Fields (CRF) to refine the concept detection of movie shots holistically, given the relationships extracted from ConceptNet and the co-occurrence information defined by normalized Google distance (NGD). We demonstrate solid improvements in performance by using mid-level concept based detectors and CRF refinement in both objective and subjective subtasks.

1. INTRODUCTION

This year, we explored several interesting possibilities in detecting the violent scenes in movies. Besides using the low-level features, we use the violence concept detectors to infer the occurrence of the violent scenes. In addition, Conditional Random Fields (CRF) are used as a refinement to improve the overall violence concept detection.

2. SYSTEM DESCRIPTION

Figure 1 shows the overview of our system framework. A diverse set of audio-visual features are extracted for training the χ^2 SVM classifiers for violent scenes detection. These low-level features include:

Dense Trajectories: The features are extracted using the method of [5]. Each trajectory is described by three features, namely histogram of oriented gradients (HOG), histogram of optical flow (HOF) and motion boundary history (MBH). Including the trajectory shape features, we have 4 features in total. Each of these features encodes some complementary information in the videos. HOG encodes the local appearance information while the local motion patterns are captured by the HOF and MBH.

SIFT: Two sparse keypoint detectors, Difference of Gaussian and Hessian Affine, are adopted to locate locally invariant image patches from video frames. This feature is then represented using the popular BoW framework, using two separate 500-d codebooks. Three spatial layers (1×1, 3×1 and 2×2) are used in the vector quantization process, producing a 8,000-dimensional feature vector by concatenating the features from both detectors.

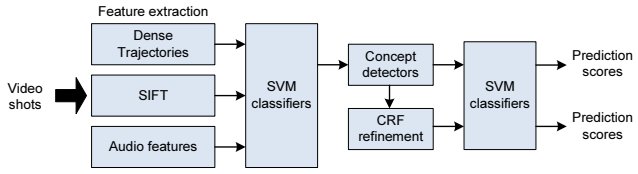

Figure 1: Framework of our system for violent scenes detection.

Audio Features: The MFCC features are densely extracted from the audio track of the videos. However, we found that MFCC is not sensitive to some audio dominant concepts, e.g. explosions and gunshots. This has inspired us to investigate the other audio features. Due to the length limit, we are not going to report on the performances of each audio feature forth. The best result is obtained with the combination of line spectral frequency (LSF), octave band signal intensity (OBSI), linear predictor coefficients (LPC), MFCC and their first and second derivatives.

We train the SVM classifiers using mid-level concept based features. These concept based features are composed of the prediction output of the violence concept detectors. The detectors are trained using the aforementioned low-level audio-visual features. Kernel-level early fusion (mean of features) is used to fuse all these low-level features. Ten violence concepts are provided by MediaEval [2], such as "fights", "explosions", "gun shots", etc. We use these 10 violence concepts to infer the other 42 extra violence concepts from the ConceptNet [4] and we train these extra violence concepts using the Youtube video clips, which are crawled using keywords and tags without human inspection. The motivation behind this is to build an event network with more diverse violence concepts. The violence-related concepts are depicted in Table 1.

We detect the occurrence of these violence concepts in the video shots and use the detection scores as the features to the SVM classifiers. Since we have collected 52 violence concepts from the ConceptNet, a graphical model can be generated to represent the violence concepts and their relationships based on the ontology of ConceptNet. We incorporate the

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

Table 1: 52 violence-related concepts inferred from ConceptNet, including the 10 violence concepts (underlined) defined by MediaEval.

accident	club	gun	person	slap
action	cold arm	gunshot	pull	stab
arm	explosion	hand	punch	stick
beat	fall	harm	punishment	victim
bleed	fight	help	push	violence
blood	fire	hit	rape	war
bomb	firearm	horror	roll	whip
bone	foot	hurt	rope	woman
break	force	kick	scream	
bullet	gang	machine gun	shock	
car chase	gore	murder	shoot	

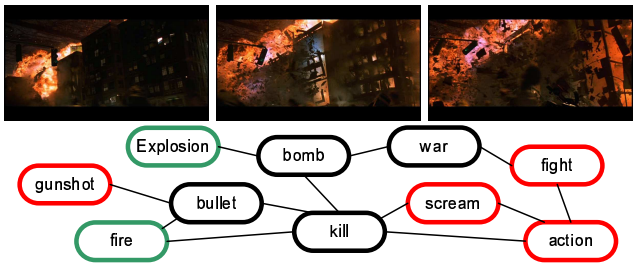

Figure 2: CRF refinement example shown in a partial event network. The retained and discarded concepts are circled in green and red respectively.

co-occurrence information of these violence concepts into the CRF for detection refinement. For example, "gory scenes" is normally co-occur with "blood" concept. Our objective is to retain certain concepts and discard the others in the event network for a particular video clip. A pairwise energy function which is making use of the detection output and also incorporating the co-occurrence statistics is proposed as follows:

$$E(X) = \sum_{v_i \in \mathcal{V}} \psi_{v_i}(x_i) + \sum_{(v_i, v_j) \in \mathcal{N}} \delta_{v_i v_j}(x_i, x_j) \qquad (1)$$

where the unary potential ψ_{v_i} is defined over the retention of concepts in the graph, based upon the classifier responses, i.e. the detection scores of SVMs. The pairwise potential $\delta_{v_i v_j}$ is defined over the co-occurrence of concepts, where normalized Google distance (NGD) [1] is adopted. Graph cut is used to minimize the energy function. The refinement process is carried out before feeding into the SVM for violent scenes detection. Figure 2 shows an example of the CRF refinement. The originally detected concepts include "action", "explosion", "fight", "fire", "gunshot" and "scream". After CRF refinement, only "explosion" and "fire" are retained. The unary potentials (the detection scores) of the discarded concepts, although beyond the thresholds, are surpassed by the pairwise potentials (the co-occurrence information) in the energy based model where the detection is considered holistically.

As we found that score smoothing [3] was very useful in improving the result performance last year, it is adopted for all the final prediction scores. The prediction scores are averaged over a three-shot windows along the timeline of each movie.

Table 2: Performance of our system for violent scenes detection.

	Objective Subtask		Subjective Subtask	
	mAP@20	mAP@100	mAP@20	mAP@100
Low-level Feat.	0.6167	0.5909	0.7223	0.6942
Concept	0.6294	0.5749	0.7688	0.7162
Concept+CRF	0.6111	0.6063	**0.8064**	0.7306
Late Fusion	**0.6509**	**0.6195**	0.7996	**0.7429**

2.1 Submitted Runs

As depicted in Table 2, we submitted four runs based on the aforementioned features, namely the low-level features (the baseline), the mid-level concept based features, the mid-level concept based features with CRF refinement and the late fusion of all the runs.

3. RESULTS AND DISCUSSION

Table 2 shows the performance[1] of our system. Each result is obtained from the mean of five repeated sets of run. It can be seen that detection using concept based features is superior to low-level features in overall. In particular, a more significant improvement is shown in the subjective subtask. The effect of CRF refinement can also be observed from the runs with CRF compared to the counterparts with no CRF. If compared to the baseline, the CRF runs in subjective subtask show a solid performance improvement of 11.6% and 5.2% in mAP@20 and mAP@100 respectively. The results indeed show the use of CRF to consider the concept detection holistically using co-occurrence information is effective. It is on the other hand shows that the structure of the event network derived from the ConceptNet is useful. Finally, the runs with late fusion benefit from having the advantages of the other runs and show the highest mAP in three out of four evaluations.

4. ACKNOWLEDGMENTS

The work described in this paper was fully sponsored by a grant from the National Natural Science Foundation of China (61272290) and was fully supported by the Shenzhen Research Institute, City University of Hong Kong.

5. REFERENCES

[1] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, Mar. 2007.

[2] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, 2013.

[3] Y.-G. Jiang, Q. Dai, C. C. Tan, X. Xue, and C.-W. Ngo. The shanghai-hongkong team at mediaeval2012: Violent scene detection using trajectory-based features. In *MediaEval 2012 Workshop*, 2012.

[4] H. Liu and P. Singh. Conceptnet – a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, Oct. 2004.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[1]The results are obtained with amended thresholds, different from the official submissions.