# RUCMM at MediaEval 2015 Affective Impact of Movies Task: Fusion of Audio and Visual Cues

Qin Jin*, Xirong Li*, Haibing Cao, Yujia Huo, Shuai Liao, Gang Yang, Jieping Xu
Multimedia Computing Lab, School of Information, Renmin University of China
Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
{qjin,xirong}@ruc.edu.cn

## ABSTRACT

This paper summarizes our efforts for the first time participation in the Violent Scene Detection subtask of the MediaEval 2015 Affective Impact of Movies Task. We build violent scene detectors using both audio and visual cues. In particular, the audio cue is represented by bag-of-audio-words with fisher vector encoding. The visual cue is exploited by extracting CNN features from video frames. The detectors are implemented using two-class linear SVM classifiers. Evaluation shows that the audio detectors and the visual detectors are comparable and complementary to each other. Among our submissions, multi-modal late fusion leads to the best performance.

## 1. INTRODUCTION

The 2015 Affective Impact of Movies Task consists of two subtasks: Induced Affect Detection and Violence Detection which we participated in for the first time. Violent scene detection (VSD) which automatically detect violent scenes in videos is a challenging task due to its large variations in video quality, content, and broad semantic meaning. Violence is defined as " *violent videos are those one would not let an 8 years old child see because of their physical violence*". MediaEval provides a common corpus and evaluation platform that encourages and enables competition and comparison among research teams. In this paper, we describe our VSD system for our first time participation in MediaEval 2015 [8]. We focus on utilizing both audio and visual cues in the video for violent scene detection. Our audio-based system uses bag-of-audio-words with fisher vector encoding, while our visual-based system uses deep features extracted by pretrained Convolutional Neural Networks (CNN) models. We combine both modalities via late fusion, and investigate two weighting strategies. One is equal weights, and the other is non-equal weights learned on a held-out subset of the development dataset.

## 2. SYSTEM DESCRIPTION

In this task, we build audio-only subsystems and visual-only subsystems. We also fuse the two modality subsystems via late fusion. The detailed description of feature representation and prediction model of each subsystem is presented in following subsections.

*Equal contribution and corresponding authors.


*MediaEval 2015 Workshop,* Sept. 14-15, 2015, Wurzen, Germany

### 2.1 Audio Feature Representation

We chunk the audio stream into small segments with some overlap (such as a 3-sec segment and 1-sec shift leading to 2-sec of overlap between adjacent segments), and empirically find that 2s segment length with 1s shift achieves the best detection accuracy. We therefore use this setup.

We use the Mel-frequency Cepstral Coefficients (MFCCs) as our fundamental frame-level feature. The MFCCs are computed over a sliding short-time window of 25ms with a 10ms shift [1]. Each 25ms frame of an audio segment is then represented as a 39-dimensional MFCC feature vector (13-dimensional MFCC + delta + delta delta). An audio segment is then represented by a set of MFCC feature vectors. Finally, we use two encoding strategies to transform this set of MFCC frames into a single fixed-dimension segment-level feature vector: Bag-of-Audio-Words (BoAW) and Fisher Vector (FV) [6].

**Bag-of-Audio-Words**: We first use an acoustic codebook to generate the segment-level feature vector. The codebook model is a common technique used in the document classification (bag-of-words) [10] and the image classification (bag-of-visual-words) [5] fields. We use the bag-of-audio-words model to represent each audio segment by assigning its low-level acoustic features (MFCCs) to a discrete set of codewords in the vocabulary (codebook), thus providing a histogram of codeword counts. The vocabulary of BoAW is learned by applying Kmeans clustering algorithm with K=4096 on the whole training dataset.

**Fisher Vector**: The Fisher Vector (FV) [6] representation can be seen as an extension of bag-of-words representation. Both the FV and BoAW are based on an intermediate representation, the audio vocabulary built in the low level feature space. The Fisher encoding uses Gaussian Mixture Models (GMM) to construct an audio word dictionary. We compute the gradient of the log likelihood with respect to the parameters of the model to represent an audio segment. The Fisher Vector is the concatenation of these partial derivatives and describes in which direction the parameters of the model should be modified to best fit the data. A GMM with 256 mixtures is used in our experiments to generate FV representation.

### 2.2 Visual Feature Representation

We consider both frame-level and video-level representations. Given a video, we uniformly extract its frames with an interval of 0.5 seconds. Subsequently, we extract CNN features from these frames. In particular, we employ two existing CNN models, i.e., the 16-layer VGGNet [7] and

GoogLeNet [9]. The feature vectors are the last fully connected layer of VGGNet, and the pool5 layer of GoogLeNet, respectively.

A video's feature vector is obtained by mean pooling the feature vectors of its frames.

## 2.3 Classification Model

For both the audio and visual systems, we train two-class linear SVM classifiers as violent scene detectors. A frame is considered as a positive training example if its video is labelled as positive with respect to the violent class. To learn from many training examples, we employ the Negative Bootstrap algorithm [3]. The algorithm takes a fixed number $N$ of positive examples and iteratively selects those negative examples, which are misclassified the most by the current classifiers. The algorithm randomly samples $10 \times N$ number of negative examples from the remaining negative examples as candidates at each iteration. An ensemble of classifiers trained in the previous iterations is used to classify each of the negative candidate examples. The top $N$ most misclassified candidates are selected and used together with the $N$ positive examples to train a new classifier. The algorithm takes several bags of positive examples and performs the training independently on each of the positive bags, resulting in multiple ensembles. They are compressed into a single vector [2], making the prediction very fast.

## 2.4 Prediction at Video Level

For detectors trained using the frame-level representations, they make prediction also at frame-level. In order to aggregate the frame-level scores to the video-level, we first apply temporal smoothing to refine scores per frame. For the visual-based system, we take the maximum response of the frames as their video score, while for the audio-based system, the video score is obtained by averaging over its frames.

We fuse the two modalities of audio and visual via simple linear fusion at the decision score level. We experiment two fusion strategies: 1) simply assigning equal fusion weights to each modality and 2) learning the optimal fusion weights via coordinate ascent [4].

# 3. EXPERIMENTS

## 3.1 Dataset

There are in total 6,144 labelled videos for development in this year's task. We split the development set randomly into two partitions, namely 1) dev-train consisting of 4,300 videos among which 190 videos are labelled as violent videos, and 2) dev-val of 1844 videos among which 82 videos are labelled as violent videos. The detectors are trained on dev-train, with hyper parameters tuned on dev-val.

## 3.2 Submitted Runs

All the runs use the previous described subsystems or fused system. We use feature name to indicate a specific system. For instance, BoAW refers to the system using the BoAW feature. Frame-level VGGNet-CNN means the system is learned from frames which are represented by VGGNet-CNN, while Video-level VGGNet-CNN means learning directly from video vectors. We submitted 5 runs:

**Run1**: Learned fusion of BoAW and FV.

**Table 1: Performance of our VSD system with varied settings. Evaluation metric: MAP.**

| System setting | dev-val | test |
|---|---|---|
| BoAW | 0.320 | – |
| FV | 0.313 | – |
| Frame-level GoogLeNet-CNN | 0.245 | – |
| Video-level GoogLeNet-CNN | 0.296 | – |
| Run1 (BoAW + FV) | 0.348 | 0.106 |
| Run2 (Frame-level VGGNet-CNN) | 0.347 | 0.118 |
| Run3 (Video-level VGGNet-CNN) | 0.308 | 0.120 |
| Run4 (Average fusion) | 0.485 | 0.216 |
| Run5 (Learned fusion) | 0.500 | 0.211 |

**Run2**: Frame-level VGGNet-CNN.
**Run3**: Video-level VGGNet-CNN.
**Run4**: Average fusion of all audio and visual runs, including BoAW, FV, Frame-level VGGNet-CNN, Video-level VGGNet-CNN, Frame-level GoogLeNet-CNN, and Video-level GoogLeNet-CNN.
**Run5**: Learned fusion of all audio and visual runs.

## 3.3 Results

The performance of our VSD system with varied settings is summarized in Table 1. We observe that fusion is always helpful. For the audio-only runs, fusion of BoAW and FV brings additional gain. Fusion of the audio and visual runs results in the best performance. Probably due to the divergence between the dev-val set and the test set, while Run2 (Frame-level VGGNet-CNN) outperforms Run3 (Video-level VGGNet-CNN) on dev-val, the latter is better on the test set. Consequently, fusion with learned weights does not yield improvement.

# 4. CONCLUSIONS

Our results show that both audio and visual modalities can perform violence detection well and the two modalities are complementary to each other and simple late fusion of two modalities leads to performance enhancement. The CNN features, although without domain-specific information engineered, can generalize well for the VSD task. In the future work, we will explore more effective fusion strategy for improving detection performance.


# Acknowledgements

This research was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the National Science Foundation of China (No. 61303184), the Beijing Natural Science Foundation (No. 4142029), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130004120006), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.


# 5. REFERENCES


[1] Q. Jin, J. Liang, X. He, G. Yang, J. Xu, and X. Li. Semantic concept annotation for user generated videos using soundtracks. In *ICMR*, 2015.

[2] X. Li and C. Snoek. Classifying tag relevance with relevant positive and negative examples. In *ACM MM*, 2013.

[3] X. Li, C. Snoek, M. Worring, D. Koelma, and A. Smeulders. Bootstrapping visual categorization with relevant negatives. *TMM*, 15(4), 2013.

[4] X. Li, C. Snoek, M. Worring, and A. Smeulders. Fusing concept detection and geo context for visual search. In *ICMR*, 2012.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[6] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3), 2013.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[8] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandrea, M. Schedl, C.-H. Demarty, and L. Chen. The mediaeval 2015 affective impact of movies task. In *MediaEval 2015 Workshop*, 2015.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[10] X. Xue and Z. Zhou. Distributional features for text categorization. *TKDE*, 21(3), 2008.