Fudan-NJUST at MediaEval 2014: Violent Scenes Detection Using Deep Neural Networks

Qi Dai[§], Zuxuan Wu[§], Yu-Gang Jiang[§], Xiangyang Xue[§], Jinhui Tang[‡]
[§]School of Computer Science, Fudan University, Shanghai
[‡]School of Computer Science and Engineering, Nanjing University of Science and Technology
{daiqi,zxwu,ygj,xyxue}@fudan.edu.cn, jinhuitang@mail.njust.edu.cn

ABSTRACT

The Violent Scenes Detection task aims at evaluating algorithms that automatically localize violent segments in both Hollywood movies and short web videos. The definition of violence is subjective: "the segments that one would not let an 8 years old child see in a movie because they contain physical violence". This is a highly challenging problem because of the strong content variations among the positive instances. In this year's evaluation, we adopted our recently proposed classification method to fuse multiple features using Deep Neural Networks (DNN). The method was named *regularized DNN*. We extracted a set of visual and audio features, which have been observed useful. We then applied the regularized DNN for feature fusion and classification. Results indicate that using multiple features is still very helpful, and more importantly, our proposed regularized DNN offers significantly better results than the popular SVM. We achieved a mean average precision of 0.63 for the main task and 0.60 for the generalization task.

1. SYSTEM DESCRIPTION

Figure 1 gives an overview of our system. In this short paper, we briefly describe each of the key components. For the task definition, data and evaluation metric, interested readers may refer to [1].

1.1 Features

Three kinds of audio-visual features were extracted, which have been observed useful in 2013.

We extracted trajectory-based motion features according to our previous work [2]. A main difference is that the new improved dense trajectories (IDT) [4] were used as the basis to replace the original dense trajectories. Four baseline features, histograms of oriented gradients (HOG), histograms of optical flow (HOF), motion boundary histograms (MBH) and trajectory shape (TrajShape) descriptors were computed. These features were encoded using the Fisher vectors (FV) with a codebook of 256 codewords. We further computed our proposed TrajMF [2] based on the HOG, HOF and MBH, by considering the motion relationships of the trajectories. As the dimension of the original TrajMF is very high, we employed the expectation-maximization principal component analysis (EM-PCA) [3] for dimension reduction, generating a 1500-dimensional representation for each feature. In total, there are seven trajectory-based features, including four baseline FV and three dimension-reduced TrajMF features. See [2] for more details.

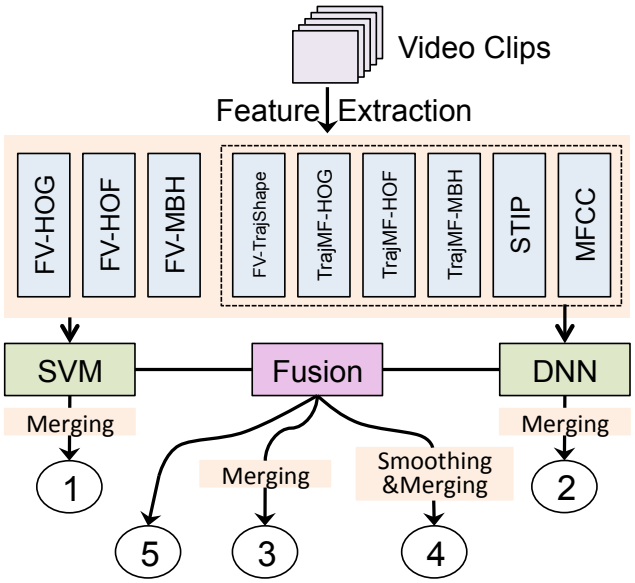

Figure 1: An overview of the key components in our system, where circled numbers indicate the 5 submitted runs.

The other two kinds of features include Space-Time Interest Points (STIP) [5] and Mel-Frequency Cepstral Coefficients (MFCC). The STIP describes the texture and motion features around local interest points, which were encoded using the bag-of-words framework with 4000 codewords. Here we randomly sampled 300k features and used k-means to generate the codebook. The MFCC is a very popular audio feature. It was extracted from every 32ms time-window with 50% overlap. The bag-of-words was also adopted to quantize the MFCC descriptors, using 4000 codewords.

1.2 Classifiers

We adopted both SVM and deep neural networks (DNN) for classification.

SVM: χ^2 kernel was adopted for the bag-of-words features (STIP and MFCC), and linear kernel was used for the others. For feature fusion, kernel-level average fusion was used for the trajectory-based features, while score-level average late fusion was adopted to combine trajectory features with STIP and MFCC.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain.

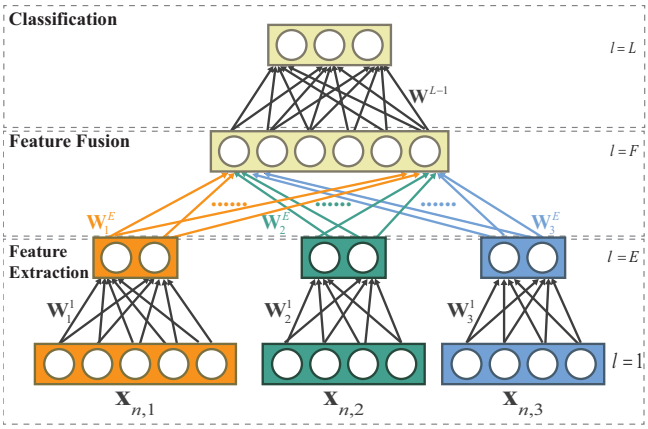

Figure 2: Illustration of the structure of our regularized DNN. Multiple features are used as the inputs, and the network transforms the features separately first, before using regularizations to explore feature relationships. The identified relationships are then utilized for improved classification performance. This figure is reprinted from [7].

DNN: We also adopted a new DNN-based classifier proposed in our recent work [6, 7]. The aforementioned fusion methods used for the SVM classifiers neglect the hidden patterns shared among the different features. To capture the relationships of distinct features, we constructed a regularized DNN for video classification. Specifically, as shown in Figure 2, in the regularized DNN, a layer of neurons were first used to perform feature abstraction separately for each input feature. After that, another layer was used for feature fusion with carefully designed structural-norm regularization on network weights, which can identify feature relationships. Finally, the fused representation was used to build a classification model in the last layer. With this special network, we are able to fuse features by considering both feature correlation and feature diversity, as well as perform classification simultaneously. See [6, 7] for more details.

1.3 Score Smoothing and Clip Merging

Temporal score smoothing has been proved to be effective as incorrect predictions on a short clip may be eliminated by considering predictions on nearby clips. All the videos were first partitioned uniformly into 3-second long clips. A smoothed prediction score of a clip is simply the average value of the scores in a three-clip window.

As we need to output segment level predictions (not on the fixed-length clip-level), we need to merge continuous clips if they are all determined to contain violence or no violence. This was done if their violence scores were all above or below a threshold, and the new score of the merged segment was set to be the average value of clips.

2. RESULTS AND DISCUSSIONS

We submitted 5 runs for official evaluation. As shown in Figure 1, Run 1 and Run 2 used SVM and DNN respectively. Run 2 did not use FV encoding of the HOG, HOF and MBH features, as the dimensionality of these three features are too high, which would jeopardize the performance of DNN when there is insufficient training data. Run 3 is the score fusion of Run 1 and Run 2. Run 4 is the score-smoothed version

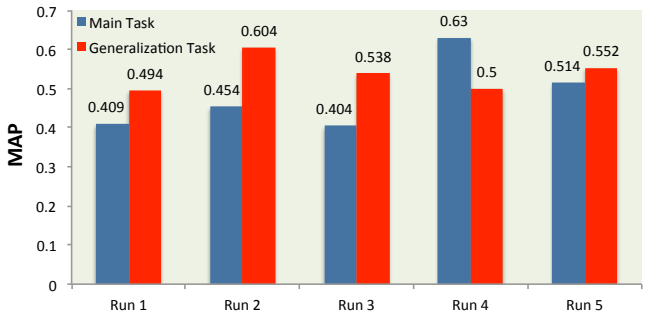

Figure 3: Performance of our 5 submitted runs on both main and generalization tasks. Note that, following this year's guideline, a specially designed MAP was used (MAP2014 [1])

of Run 3 (smoothing was performed before merging), while Run 5 is the direct fusion of SVM and DNN without using any smoothing and merging functions.

The official results are summarized in Figure 3. We see that, although some features were not used in DNN, the performance of DNN (Run 2) is still significantly better than SVM. This clearly confirms the effectiveness of deep networks. Directly fusing DNN and SVM incurs a small performance drop (Run 3). This may be due to the sub-optimal parameters used in the fusion process. Another fusion setting (Run 5) without using score merging improves the main task performance but still hurts the result of the generalization task, showing that DNN has better generalization capability than the SVM, and thus fusing SVM with DNN will always degrade the performance of the generalization task. Finally, the results of Run 4 indicate that both smoothing and merging are useful for the main task. It is not surprising that smoothing does not work for the generalization task, because, compared with the long movies used in the main task, the test clips are short and are relatively temporally more consistent.

Acknowledgements

This work was supported in part by a National 863 Program (#2014AA015101), the National Natural Science Foundation of China (#61201387), and the Science and Technology Commission of Shanghai Municipality (#13PJ1400400, #13511504503, #12511501602).

3. REFERENCES

[1] M. Sjöberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval 2014 Workshop*, Barcelona, Spain, Oct 16-17, 2014.

[2] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*, 2012.

[3] S. Roweis. EM Algorithms for PCA and SPCA. *NIPS*, 1998.

[4] H. Wang, C. Schmid. Action Recognition With Improved Trajectories. In *ICCV*, 2013.

[5] I. Laptev. On space-time interest points. *IJCV*, 64:107–123, 2005.

[6] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, X. Xue. Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification. In *ACM MM*, 2014.

[7] J. Tu, Z. Wu, Q. Dai, Y.-G. Jiang, X. Xue. Challenge Huawei Challenge: Fusing Multimodal Features with Deep Neural Networks for Mobile Video Annotation. In *ICME*, 2014.