Fudan Team at TRECVID 2014: Multimedia Event Detection

Zuxuan Wu, Rui-Wei Zhao

School of Computer Science, Fudan University, Shanghai, China

Abstract. In this notebook paper, we describe the submissions of Fudan Team to the Multimedia Event Detection task for TRECVID 2014. Our system exploits popular low-level descriptors to capture visual appearance, motion and audio information from a video clip. In addition, it also incorporates the high-level semantic feature generated by a Convolutional Neural Network pre-trained on ImageNet. We performed classification with SVMs. We submitted results for the full MED14 evaluation in two (010Ex and 100Ex) training conditions.

1 System Overview

For TRECVID 2014 [1], we participated in the Multimedia Event Detection (MED) task. Fig. 1 presents the framework of our system. We first extract various low-level visual appearance, motion and audio features, as well as the high-level semantic feature. Then both Fisher Vector (FV) and Bag-of-Words (BoW) are adopted to produce quantized feature representations. SVMs are utilized to classify the features. Finally, the output scores from different classifiers are combined to produce a final prediction with fusion parameters estimated on the development set.

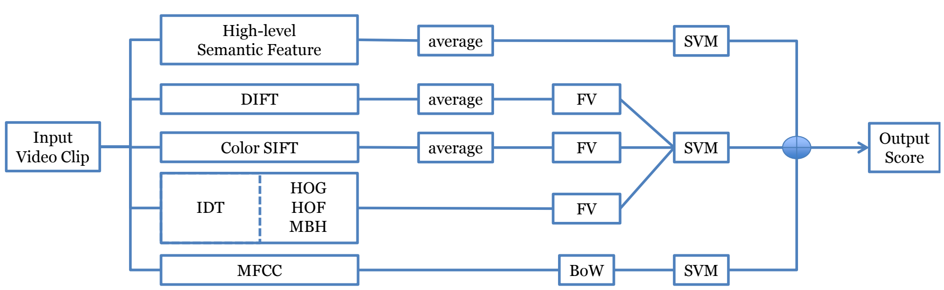

Fig. 1. The framework of our system for processing a video clip.

2 System Components

In this section, we elaborate the technical components of our system. First, we describe the adopted features as well as their corresponding encoding strategies. Then we introduce the classifiers for model training and different fusion approaches.

2.1 Feature Representation

- **Motion Features**: Motion information plays a significant role for event detection. In our system, motion is captured using the state-of-the-art improved dense trajectory features [2], which exhibits top-notch performance on action recognition tasks. Along with the densely extracted trajectories, three features are computed: HOG, HOF, and MBH. These features are further quantized respectively using the FV representation with the vocabulary size being 256.

- **Appearance Features**: To capture static visual appearance information, we adopt the dense SIFT (DIFT) [3] feature and the Color SIFT [4] feature. Here, given a video frame, we extract these two appearance features and then quantize them into FV representations with a codebook of 256 codewords separately. Then, frame-level features are averaged to generate a video-level feature representation.

- **MFCC Audio Feature**: In addition to the above visual features, audio features can provide complementary clues. For this, we adopt the well-known Mel-Frequency Cepstral Coefficients (MFCC). It is first computed over each 32ms time-window (with 16ms overlap) of the soundtrack and then all the descriptors are quantized into a single BoW feature representation.

- **High-level Semantic Feature**: We also extract the high-level semantic feature with a Convolutional Neural Network pre-trained on the ImageNet 2010 Challenge data, which consists of 1.2 million images totaling 1,000 concepts. For each key frame in a given video, we obtain a 1,000-d concept score with the trained model. Then frame-level scores are then averaged to generate a video-level concept feature vector for further classification.

2.2 Classification and Fusion

To train event detection models, we employ two different types of classifiers in our system:

- Linear SVMs: To enhance classification performance, we first perform early fusion with the appearance feature and motion feature by concatenating them into a long vector. Since the concatenated vector is discriminative enough in the high-dimensional space, we adopt linear SVMs with C fixed to 100 to train the model.
- Non-linear SVMs: We first map features with BoW representation and the high-level semantic features into χ^2-kernel separately. Then, we train two independent classifiers.

With multiple classifiers, each video clip is accordingly associated with multiple output scores, which are then fused to compute the final prediction.

3 Experiments

In this section, we present experimental results obtained on the *development* set of this year and report our official results on the MED14-Test dataset. Table 1 presents the performance of individual features and their combinations under the 010Ex training condition on the development set. We can see that visual features outperform the high-level semantic feature and the MFCC feature significantly. Table 1 also demonstrates that the fusion of multiple features promotes the overall performance. More specifically, combining visual features with the high-level semantic feature gives a 2.55% performance gain. In addition, the fusion of all features achieves the best performance.

Features	mAP
Visual (Appearance + Motion)	15.12%
High-level Semantic Feature	7.18%
MFCC Audio	2.25%
Visual + High-level Semantic Feature	17.67%
Visual + High-level Semantic Feature + MFCC Audio	18.92%

Table 1. Performance of individual features and their combinations.

Our official submissions for the full MED14 evaluation include the 010Ex and 100Ex training conditions. For the Pre-Specified task, we achieved a 10.7% mAP (010Ex) and a 22.1% mAP (100Ex); for the Ad-Hoc task, we achieved a 7.4% mAP (010Ex) and a 15.6% mAP (100Ex). Notice that although we discovered the high-level semantic feature could be extremely helpful on the development set, unfortunately we found a bug in the extraction phase of this feature on the MED14-Test dataset. Regretfully, our official submissions did not take advantage of the powerful feature, with which we could obtain better results.

References

1. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quenot, G.: Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2014. (2014)
2. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision. (2013)
3. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time visual concept classification. IEEE Transactions on Multimedia (2010)
4. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)