The 2011 SESAME Multimedia Event Detection (MED) System

SRI International	Murat Akbacak, Robert C. Bolles, J. Brian Burns, Mark Eliot, Aaron Heller, James A. Herson, Gregory K. Myers, Ramesh Nallapati, Eric Yeh
University of Amsterdam	Dennis C. Koelma, Xirong Li, Masoud Mazloom, Koen E.A. van de Sande, Arnold W.M. Smeulders, Cees G.M. Snoek
University of Southern California	Sung Chun Lee, Ram Nevatia, Pramod Sharma, Chen Sun, Remi Trichet

ABSTRACT

The SESAME team submitted four MED-11 runs which combined video content extraction results consisting of visual features, video OCR results, and motion features. The primary run and one of the secondary runs used two different methods of fusing visual features and OCR results; a third run combined visual features and motion features; and a fourth run combined visual features, OCR results, and motion features. Results were combined using rank-based fusion and weighted averages. We found that rank-based fusion of visual feature results and video OCR results (the primary run) had the best performance of the four runs. The initial performance of the runs with motion features, which were computed around keyframes, was poor, but a subsequent experiment showed that motion features can indeed contribute to improved performance.

1 INTRODUCTION

The SESAME team consists of SRI International (SRI), the University of Amsterdam (UvA), and the University of Southern California (USC). We submitted four MED-11[1] runs which combined video content extraction results from each team member:

- Visual features from UvA
- Video OCR from SRI
- Motion features from USC

The primary run and one of the secondary runs used two different methods of fusing visual features and OCR results; a third run combined visual features and motion features; and a fourth run combined visual features, OCR results, and motion features.

In this paper, we describe the content extraction methods (section 2),fusion methods for the four runs (section 3),threshold selection methods (section 4), and the experimental results (section 5).

2 CONTENT EXTRACTION METHODS

2.1 Visual Features

We extracted visual features by dividing each video clip into shots and selecting key frames equally spaced throughout each shot. The visual features consist of vector-quantized Harris-Laplace and Dense-Sampled SIFT, OpponentSIFT, and RGB-SIFT descriptors, with spatial pyramids and a Support Vector Machine [2]. These were augmented with features resembling Fisher Vectors [3]. Event detectors were trained by manually annotating a selection of single frames from the event kits. In each video, event detection is performed on every key frame, plus six extra I-frames around each keyframe. The score of the video is the maximum score of all the frames analyzed within that video.

2.2 Video OCR

SRI's video OCR software recognized and extracted text from MED-11 video imagery. This software recognizes both overlay text, such as captions that appear on broadcast news programs, and in-scene text on signs or vehicles [4]. It currently detects and reads printed text only.

The process detects candidate text lines in each frame, separates them from the background, and tracks them from frame to frame. When a tracked text line is no longer detected in the scene, samples of that text line are binarized, preprocessed, and sent to an OCR engine. If the text line is being viewed from an oblique angle in the scene, the process rectifies it by removing perspective distortion before binarization. For each tracked text line, the recognized characters from multiple frames are collectively analyzed to form a single recognition result.

The recognized text was filtered using several criteria. To minimize the chance that the recognized text was generated from an image pattern and not text in the image (false text), recognized text was included only if it appeared in the video for at least one second. An on-line dictionary was used to filter out any non-English content. Because longer words are less likely to be false text, only words with a minimum number of characters were retained, and English stop words were removed.

Event models consisted of the log frequency counts of individual words. The event models were trained on text extracted from video exemplars in the Event Kit. For each test video, a Bag-of-Words was formed using log-adjusted frequencies of the words observed in the video. We used cosine and Jaccard similarity measures to match the event models to text found in each test video.

2.3 Motion Features

We used an approach based on Spatio-Temporal Interest Points (STIPs)[5]. Corner-like interest points were found in the spatio-temporal volume and local gradient, and visual features are computed around these points. The STIP feature set, which was obtained from the SRI Aurora team, contained histograms of gradient (HOG) and flow (HOF) descriptors. The features were extracted from the MED-11 video clips and clustered into 10,000 motion code words.

Our aim was to characterize the motion around representative keyframes of the event (as determined by the manual annotation process for visual features). We formed histograms with the STIP features falling within a 100-frame window centered on each keyframe. For each event, we trained a Support Vector Machine (SVM) classifier using a Chi-square kernel on the STIP histograms in videos from the Event Kit. Testing was performed by applying the classifier to the keyframes in each test video (with up to six extra I-frames per shot) and selecting the maximum score over the frames analyzed within that video.

3 FUSION METHODS

Only a small fraction of MED-11 videos contain readable text. Based on the results we observed for the training events, we derived two procedures for fusing the results from the visual features and the video OCR. The first method fused the results by interleaving the five top-ranked video OCR results among the five top-ranked results for the visual features. The second procedure computed a weighted average of the confidence scores (only for videos which produced video OCR output). The first procedure was used in the primary run, and the second procedure was used in the second run.

The third run fused the results of the visual features and the motion features by averaging their confidence scores. The fourth run fused the confidence scores of the third run with the scores of the video OCR using the weighted average.

4 AUTOMATIC THRESHOLD SELECTION

Our goal was to design a threshold selection model that would minimize the Normalized Detection Cost (NDC) [1] on unseen test data. The NDC is a weighted combination of missed detection and false alarm probabilities. A single fixed threshold would not minimize mean NDC over all training events; therefore, we developed a model that would adapt to each event.

To develop the model, we found the optimal threshold for each training event by walking through the entire ranked list to find the value with the minimum NDC. Optimal thresholds for each of the five training events occurred between the 1% and 2% rank positions. We then formed a linear regression model that maps confidence scores at the 1% and 2% rank positions to the threshold value that corresponds with the minimum NDC. We then applied this model to select the threshold for each of the test events.

5 EXPERIMENTAL RESULTS AND CONCLUSIONS

Four runs were submitted to the MED-11 evaluation:

- Run 1: Visual feature results combined with video OCR results using a rank-based fusion of the confidence scores
- Run 2: Visual feature results combined with bag-of-words video OCR results using a weighted average of confidence scores
- Run 3: Visual feature results combined with the results of key-interval-based STIPs motion features using a weighted average of confidence scores
- Run 4: Run #3 combined with video OCR results using a weighted average of confidence scores

The results are shown in Figures 1 through 4.

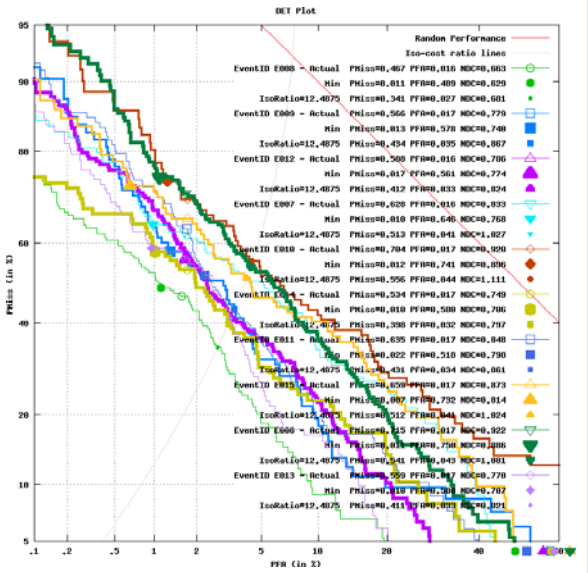

Figure 1. Results from run #1.

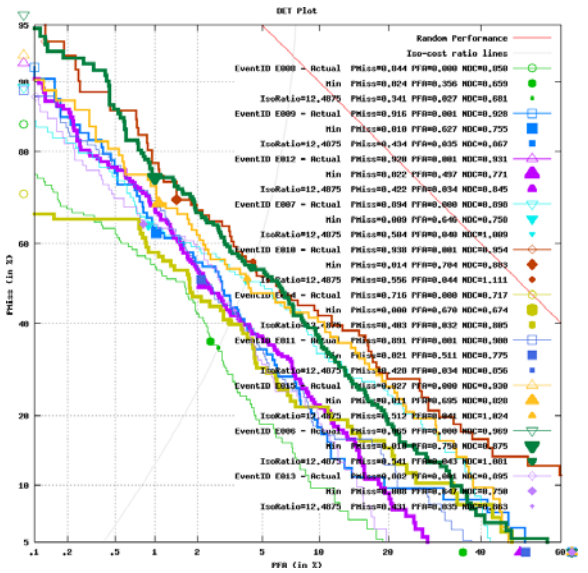

Figure 2. Results from run #2.

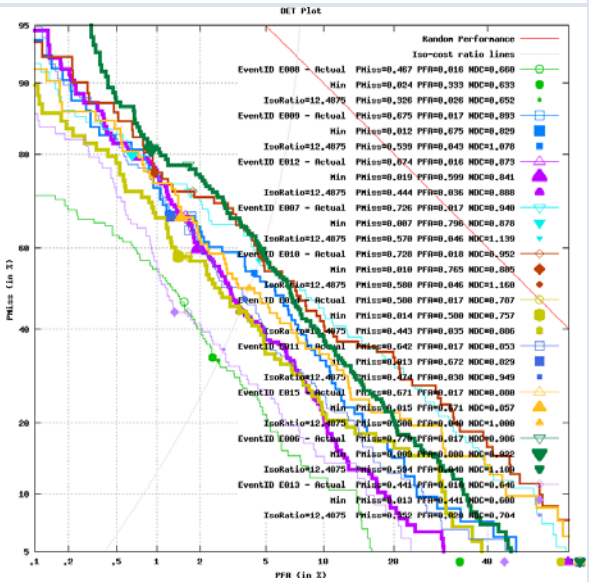

Figure 3. Results from run #3.

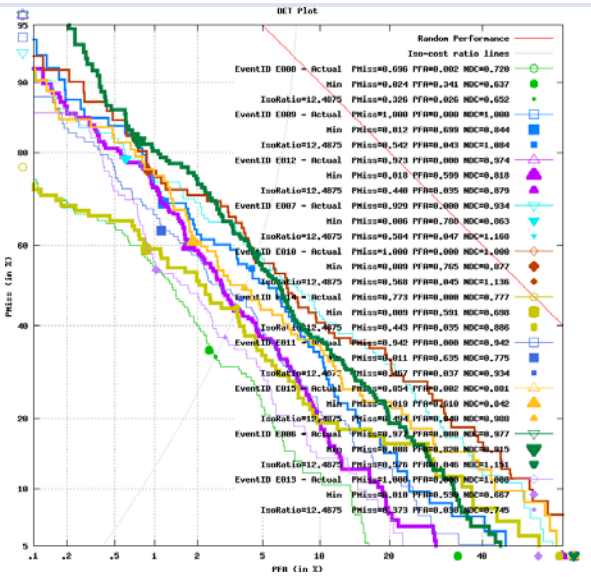

Figure 4. Results from run #4.

From these results, we draw the following conclusions:

- The performance of rank-based fusion of visual feature results and video OCR results (Run #1) had the best performance of the four runs. Additional tests showed that its performance was greater than the performance of the visual feature results alone, thereby confirming that including the video OCR results does indeed improve performance.
- Combining video OCR confidence scores with confidence scores of the visual features by a weighted average (Runs #2 and #4) had a negligible effect on performance. Therefore, we conclude that this method of fusion was less effective than the ranked-based fusion of Run #1.

- Contrary to our expectations, the performance of combining visual features with motion features(Run #3) was poorer for all events except for the event “Parkour.”
- Our method of selecting a decision threshold skewed the detections towards extremely low false alarm rates at the expense of increased miss rates. At first, it appeared that we optimized the wrong criterion: instead of the minimum NDC, we should have optimized the NDC score at the TER line. In subsequent experiments, we confirmed that using the latter criterion resulted in operating points that were much closer to the TER line.

These results prompted post-evaluation analysis to better understand the performance:

- We found a high degree of correlation between detections from the visual-feature-based classifier and the motion-based classifier. This is perhaps because the motion-based classifier had used a histogram of STIP features falling within a 100-frame window centered on each keyframe found by the visual-feature-based classifier. Therefore, in post-evaluation experiments, histograms of STIP features were computed across the entire video clip instead of only within the window around each keyframe.
- We experimented with additional methods to perform the late-stage classifier fusion: Support Vector Machines with multiple kernel types, L2-penalized Logistic Regression, and an Expectation Maximization (EM) based framework. We found that EM performed the best. We believe EM is attractive because, unlike in a meta-classifier, none of the learned mixture weights can be negative.

Figure 5 compares the performance of EM-based fusion of visual feature results and results from motion features computed across the entire video clip (red curve) with visual feature results only (green curve). Each curve represents the average of the curves for the 10 test events (to make it easier to visualize the comparison of the two runs). The curves show that the fusion results out-perform the results with visual features only at all operating points, which demonstrates that motion features can indeed contribute to improved performance.

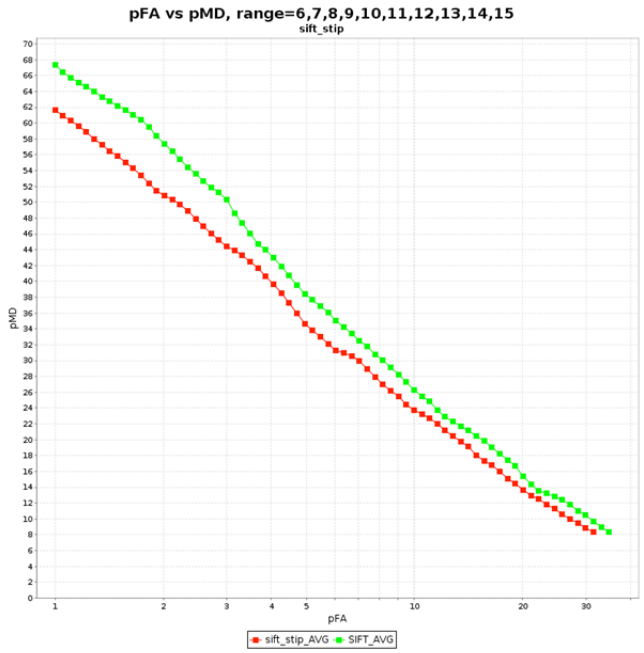

Figure5. Comparison of the Fusion model of visual and motion features with the visual-features-only model on the DEVO set in terms of the DET curves averaged over the 10 test events. The fusion results out-perform the results with visual features only at all operating points.

6 ACKNOWLEDGEMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center, contract number D11PC0067.The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7 REFERENCES

[1] Smeaton, A. F., P. Over, G. Quénot, et al., "TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics." http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/tv10overview.pdf

[2] Snoek, C.G., M. Koen, E. A. van de Sande, Xirong Li, Masoud Mazloom, Y.-G. Jiang, Dennis C. Koelma, and Arnold W. M. Smeulders, "The Media Mill TRECVID 2011 Semantic Video Search Engine," in *Proceedings of the 9th TRECVID Workshop*, Gaithersburg, USA, 2011.

[3] Perronnin, Florent, Jorge Sanchez, and Thomas Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in ECCV, 2010.

[4] Myers, G., R. Bolles, Q.-T. Luong, J. Herson, H. Aradhye, "Rectification and recognition of text in 3-D scenes," *International Journal on Document Analysis and Recognition*, Vol. 7, No. 2-3, July 2005, pp. 147-158.

[5] Laptev, I., "On space-time interest points." *International Journal of Computer Vision*, *64* (2–3), pp. 107–123, (2005).