Bilkent University at TRECVID 2006

S. Aksoy, P. Duygulu,
G. Akçay, E. Ataer, M. Baştan, T. Can, Ö. Cavuş, E. Doğrusöz, D. Gökalp,
A. Akaydın, L. Akoğlu, P. Angın, G. Cinbiş, T. Gür, M. Ünlü

RETINA Vision and Learning Group
Department of Computer Engineering
Bilkent University
Bilkent, 06800, Ankara, Turkey

Abstract

We describe our third participation, that includes one high-level feature extraction run, and two manual and one interactive search runs, to the TRECVID video retrieval evaluation. All of these runs have used a system trained on the common development collection. Only visual and textual information were used where visual information consisted of color, texture and edge-based low-level features and textual information consisted of the speech transcript provided in the collection.

1 Introduction

This is the third participation of the RETINA Vision and Learning Group at Bilkent University to TRECVID. The team that participated to TRECVID included six undergraduate students and seven graduate students supervised by two faculty members. We have developed a system for automatic classification and indexing of video archives as part of undergraduate research projects. This paper summarizes the approaches we have taken in one high-level feature extraction run, and two manual and one interactive search runs we have submitted this year.

2 Preprocessing

In all of the runs, we have used the shot boundaries, keyframes, speech transcripts and manual annotation provided with the TRECVID 2005 and 2006 data.

The speech transcripts are in the free text form and require preprocessing. First, we use a part of speech tagger to extract nouns which are expected to correspond to object names. Then, we apply a stemmer and remove the stop words and also the least frequent words to obtain a set of descriptive words. Finally, we make use of WordNet to construct a hierarchy to extend the word set for each keyframe. For example, even when the word "sport" is not included in the speech transcript, if any kind of sport such as "basketball" or "soccer" appear in the speech transcript, that keyframe is also associated with the word "sport" to allow more flexibility during classification and retrieval.

We model spatial content of images using grids. The low-level features based on color, texture and edge are computed individually on each grid cell of a non-overlapping partitioning of 352×240 video frames into 5 rows and 7 columns. Each resulting grid cell is associated with the statistics

(mean and standard deviation) of RGB, HSV and LUV values of the corresponding pixels as the color features and the statistics of the Gabor wavelet responses of the pixels at 3 different scales and 4 different orientations as the texture features. Histograms of the gradient orientation values of the Canny edge detector outputs are used as the edge features. Orientation values are divided into bins with increments of 45 degrees and an extra bin is used to store the number of non-edge pixels.

This process results in 5 feature vectors for each grid cell with the following lengths: 6 for each of RGB, HSV and LUV statistics, 24 for Gabor statistics, and 9 for edge orientation histograms. Individual components of each feature vector are also normalized to unit variance to approximately equalize ranges of the features and make them have approximately the same effect in the computation of similarity.

3 High-Level Feature Extraction

We have developed a generic classifier that uses the low-level features described with the k-nearest neighbor rule. First, we performed experiments using different feature combinations and different k values on the TRECVID 2005 data. We have empirically decided to use tiled HSV histograms and Canny-based edge orientation histogram features. The k value was also chosen as 51. The examples provided with the common annotation were used to train the classifier for all high-level features.

We have observed that the resulting classifier was particularly effective for features such as weather, sport, studio setting, computer screen, and map. Evaluation of the classifier on TRECVID 2006 data also reflected this observation.

4 Search

The baseline manual run used processed ASR outputs. This run (Bilkent2) was particularly effective for topics that involve specific keywords such as "president", "Bush", "Chaney", "Rice", etc.

The second manual run used only low-level visual features that were also used in the high-level feature extraction task. The shots were sorted according to their distance to the query shot using these low-level features. This run (Bilkent1) was very effective for topics that have specific color content such as "soccer", "helicopter" (gray vehicle on blue/gray sky) and "flame". An example query is shown in Figure 1.

Finally, the interactive run (Bilkent3) that used both ASR keywords and low-level visual features combined the advantages of both features and gave satisfactory results for many more topics. Example queries are shown in Figures 2 and 3. These results show the potential of multi-modal analysis for future research.

5 Conclusions

Our participation to TRECVID consisted of one high-level feature extraction run, and two manual and one interactive search runs this year. We are currently working on extending our system with new low-level features, classifiers and novel methods for their multi-modal fusion.

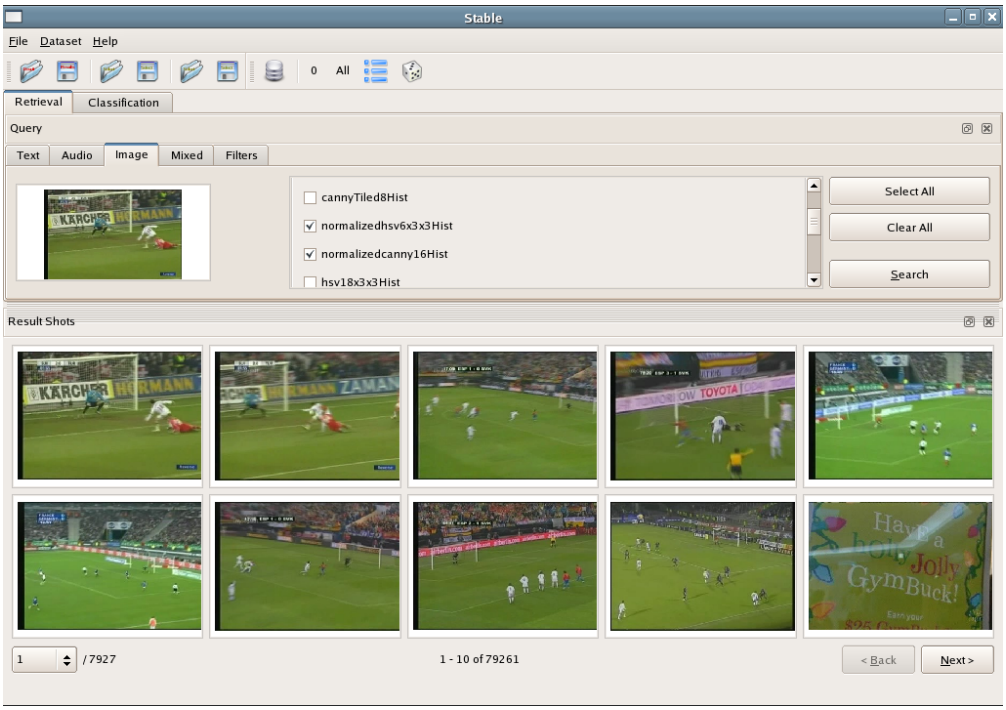

Figure 1: Manual search results using low-level color and edge-based features for a "soccer" query.

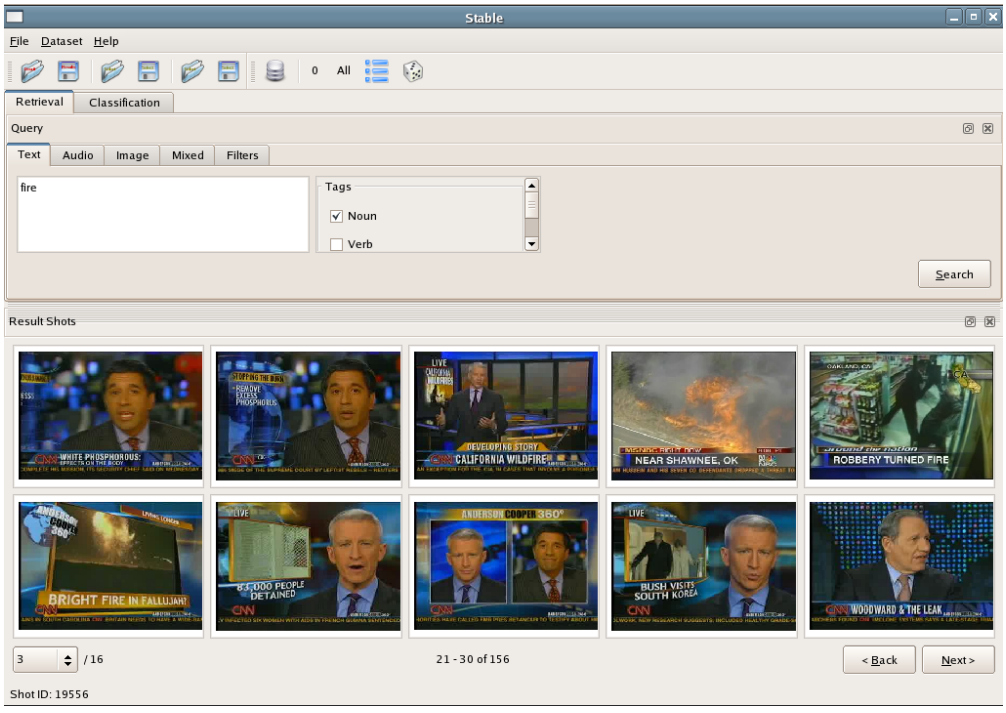

Figure 2: Interactive search results using the baseline (ASR-based) system for a "fire/flame" query.

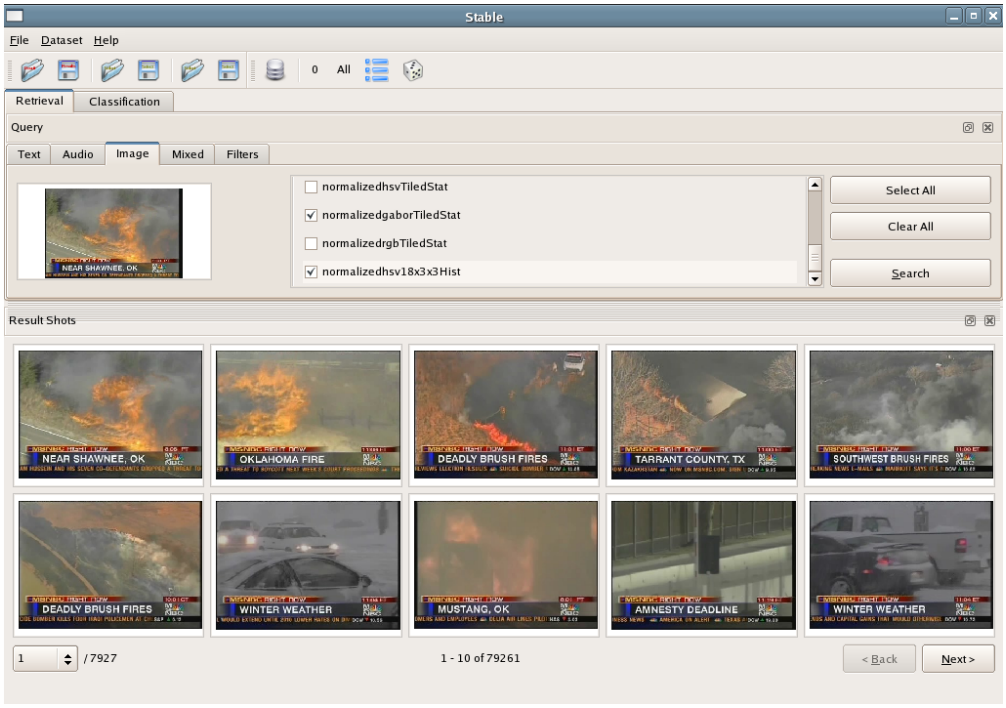

Figure 3: Interactive search results using low-level features for a "fire/flame" query.