Bilkent University at TRECVID 2005

S. Aksoy, A. Avcı, E. Balçık, Ö. Çavuş, P. Duygulu, Z. Karaman, P. Kavak,
C. Kaynak, E. Küçükayvaz, Ç. Öcalan, P. Yıldız

Department of Computer Engineering
Bilkent University
Bilkent, 06800, Ankara, Turkey

Abstract

We describe our second-time participation, that includes one high-level feature extraction run, and three 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. With the experience gained with our second-time participation, we are in the process of building a system for automatic classification and indexing of video archives.

1 Introduction

This is the second-time participation of the RETINA Vision and Learning Group at Bilkent University to TRECVID. The team that participated to TRECVID included eight undergraduate students and one graduate student supervised by two faculty members. We are in the process of building 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 three 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 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.

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 could submit results only for the "people walking/running" feature on time (run ID: Bilkent). Our moving person detector first performs face detection on each keyframe and eliminates the shots that do not contain any faces. Then, a seven frame window is used around each keyframe to check whether there is a movement in the area below each face. Simple frame subtraction is used for detecting motion.

4 Search

We have performed three manual and one interactive search run. In our retrieval scenario, the user starts a search session by typing one or more words as the text query. An initial set of shots is returned by searching for these words in the speech transcripts processed as in Section 2. We use this set representing the baseline as our first manual run (run ID: Bilkent1). The second manual run (run ID: Bilkent2) uses combination of the result set from keyword-based search and face detection. The combination is done by keeping shots that are in both sets. This combination is used to perform queries that involve people.

The third manual run (run ID: Bilkent3) uses different color features. The user starts manual search using an example shot as the visual query, and a content-based search is performed on the keyframes representing all shots in the database. The similarity between images is measured by appending all feature vectors for all grid cells and computing Euclidean distances between the resulting vectors. This query assumes that each feature in each grid cell has equal contribution to the computation of similarity. This technique gave results that are better than the median of average precision in graphic map (topic ID: 155), helicopter (topic ID: 158), palm trees (topic ID: 166), and soccer (topic ID: 171) queries.

Finally, the interactive run (run ID: Bilkent4) uses different visual features and relevance feedback. Upon being presented the results of a manual run (as in the third run described above), the user labels some of these shots as relevant (positive) and irrelevant (negative). The feedback information is incorporated into the search in terms of iterative retrievals by modifying the contributions of different feature vectors from different grid cells in the overall similarity computation. These modifications are done via assigning a weight to each feature vector for each grid cell and updating these weights in subsequent iterations. As shown in Figure 1, there is a weight w_{ijk} assigned to the k'th feature vector of the cell located at the i'th row and j'th column of the grid where $i = 1, \ldots, 5$, $j = 1, \ldots, 7$ and $k = 1, \ldots, 5$. Given two images, first, distances d_{ijk} between the corresponding feature vectors and grid cells are computed, and then, these distances are combined as the overall (dis)similarity value

$$d = \sum_i \sum_j \sum_k w_{ijk} d_{ijk}. \tag{1}$$

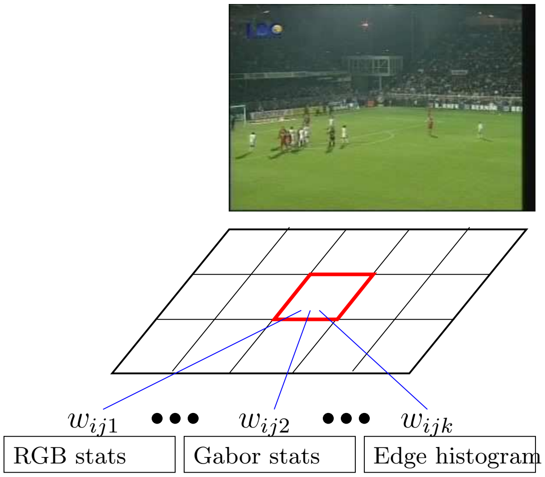


Figure 1: A 3 × 5 grid layout and the corresponding weights for a particular grid cell. A 5 × 7 layout and 5 feature vectors are used in the experiments.

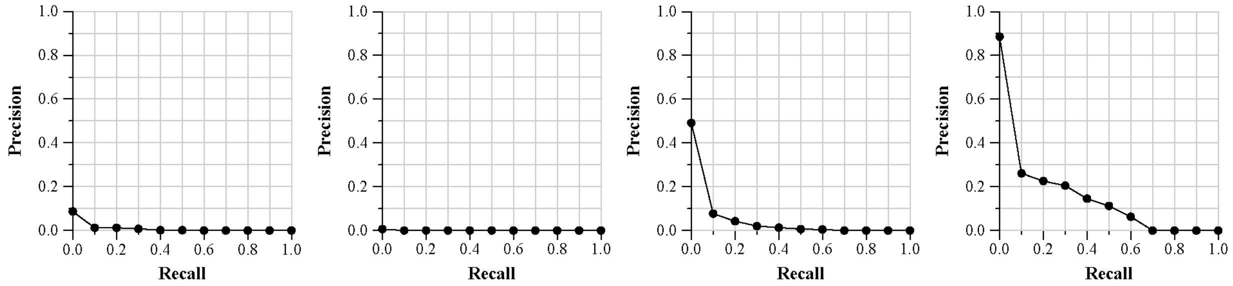
We do not use separate weights for individual feature components because of the large number of parameters required to be estimated at each iteration and the potentially low number of feedback examples and the corresponding small sample issues.

The weights are assigned uniformly in the first iteration. In earlier work [1], we have used the following assumption to compute weights for individual feature components: for a feature to be good, its variance among all the images in the database should be large, its variance among the irrelevant images should also be large, but its variance among the relevant images should be small. Here, we use a similar approach to compute the weights for different feature vector distances and grid cells. Given the positive and negative examples, for a feature vector in a particular grid cell being significant for a particular query, the distances for the corresponding vectors for relevant images must usually be similar (hence, a small variance), but the distances between the vectors for relevant images and irrelevant images must usually be different (hence, a large variance). Therefore, the weights are computed using the ratio of the standard deviation of the distances between relevant and irrelevant images to the standard deviation of the distances between relevant images.

The resulting weights represent features and the particular grid locations that are significant for a particular query session. For example, texture features are more important for representing the crowd located at the upper portion of the image in Figure 1 whereas color features are more important for representing the soccer field in the lower half of the image.

In the experiments done using the development data, among all the combinations, using color features (RGB, HSV, LUV) and Gabor features gave the highest average precision. The most significant improvement was obtained for sports-based queries where the accuracy was close to the median of average precision for tennis (topic ID: 156) and soccer (topic ID: 171). Therefore, we have done more detailed analysis of sports-based queries. Adding edge orientation features did not improve the performance for these topics because there was no significant edge structure in the categories of interest. We expect that these features will be useful for some other categories.

Apart from sports-based queries, searches for other topics did not return enough relevant shots that are worth running feedback on. However, when feedback was used, results were always improved over the manual case, i.e., the first iteration. Average precision vs. recall for the search runs are given in Figure 2. More details about the interactive run can be found in [2].



(a) Bilkent1

(b) Bilkent2

(c) Bilkent3

(d) Bilkent4

Figure 2: Precision vs. recall for search runs. Bilkent1: manual baseline run that uses only speech transcripts; Bilkent2: manual run that uses speech transcripts and face detection; Bilkent3: manual run that uses different color features; Bilkent4: interactive run that uses different visual features and weight-based relevance feedback.

5 Conclusions

Our participation to TRECVID consisted of one high-level feature extraction run, and three manual and one interactive search runs this year. The high-level feature extraction run used face detection and simple motion estimation based on frame subtraction to detect “people walking/running”. The search runs used keyword-based querying of speech transcripts, face detection, color, texture and edge features extracted using grid-based partitioning of keyframes, and relevance feedback-based combination of these features for user interaction. We are in the process of building a system for automatic classification and indexing of video archives by integrating visual, aural and textual information.

References

[1] S. Aksoy, R. M. Haralick, F. A. Cheikh, and M. Gabbouj. A weighted distance approach to relevance feedback. In *Proceedings of 15th IAPR International Conference on Pattern Recognition*, volume IV, pages 812–815, Barcelona, Spain, September 2000.

[2] S. Aksoy, O. Cavus. A relevance feedback technique for multimodal retrieval of news videos. In *Proceedings of EUROCON*, Belgrade, Serbia & Montenegro, November 21-24, 2005.