A Relevance Feedback Technique for Multimodal Retrieval of News Videos

Selim Aksoy, *Member, IEEE* and Özge Çavuş

***Abstract*— Content-based retrieval in news video databases has become an important task with the availability of large quantities of data in both public and proprietary archives. We describe a relevance feedback technique that captures the significance of different features at different spatial locations in an image. Spatial content is modeled by partitioning images into non-overlapping grid cells. Contributions of different features at different locations are modeled using weights defined for each feature in each grid cell. These weights are iteratively updated based on user's feedback in terms of positive and negative labeling of retrieval results. Given this labeling, the weight updating scheme uses the ratios of standard deviations of the distances between relevant and irrelevant images to the standard deviations of the distances between relevant images. The proposed technique is quantitatively and qualitatively evaluated using shots related to several sports from the news video collection of the TRECVID video retrieval evaluation where the weights could capture relative contributions of different features and spatial locations.**

***Index Terms*— Relevance feedback, video retrieval, TRECVID, news videos, sports videos**

I. Introduction

With the recent developments in technology, large quantities of multimedia data have become available in both public and proprietary archives. News videos, consisting of visual, textual and audio data, are important multimedia sources because of their rich content and high social impact [7]. Designing systems for content-based indexing, analysis and retrieval of these archives have become a challenging research area.

Content-based image and video retrieval techniques [9], [12] start with extracting low-level features from images and compute similarity between them using distances between feature vectors. Initial work on content-based retrieval focused on extracting color and texture features globally from an entire image. More recent work extended content extraction to region-based analysis where feature vectors are computed from segmented regions and similarity is evaluated between individual regions [3], [6]. Segmentation is performed either using semi-supervised methods that involve human involvement or using unsupervised techniques that automatically find region boundaries.

Even though successful applications of globally extracted feature representations have been shown on limited data sets, a significant portion of the content found in the news archives contains recordings taken in both outdoor and indoor scenes that are completely uncontrolled with respect to factors such as illumination, pose, location, scale, occlusion, etc. Therefore, the object variability and background complexity in these images cannot be modeled using the limited expressive power of global features. On the other hand, we have observed that many popular automatic segmentation techniques [3], [11], that have been commonly used with natural images with a small number of homogeneous regions (such as the ones in the Corel data set), do not perform well on the low-resolution news data because of both high object variability and low data quality.

In this paper, we describe a relevance feedback technique to capture the significance of different features at particular spatial locations in an image. We model the spatial content by partitioning images into non-overlapping grid cells. The low-level features such as color, texture and edge are computed individually for each cell and a weighted combination of distances based on these features is used for computing similarity between image pairs. The weights defined for each feature in each grid cell are used to capture the contribution of that feature for that particular spatial partition. User's relevance feedback in terms of positive and negative labeling of retrieval results is used to update these weights in iterative retrievals.

Earlier relevance feedback techniques include query point movement [6] borrowed from document retrieval literature, and weighting individual features and updating these weights heuristically in iterative retrievals [10], [2]. More recently, optimization-based techniques that try to compute optimum weights [8] or feature transformations [5], and support vector machine-based techniques [6], [4] that use positive and negative feedback examples to learn to classify database images have become popular. However, optimization-based techniques are not applicable when the number of feedback examples are small [8] and support vector machine-based techniques may also face stability problems due to small sample issues [13].

The proposed technique uses relevance feedback for weight updating that is based on the ratios of standard deviations of the distances between relevant and irrelevant images to the standard deviations of the distances between relevant images. Effectiveness of the proposed approach is evaluated using shots related to several sports from the news video collection of the TRECVID video retrieval evaluation [7].

The rest of the paper is organized as follows. Image representation in terms of low-level features on a spatial grid layout is described in Section II. The retrieval scenario using

This work was supported by the TUBITAK Grant 104E077 and the European Commission COST 292 Action.

S. Aksoy is with Bilkent University, Department of Computer Engineering, Ankara, 06800, Turkey. Phone: +90-312-2903405; Fax: +90-312-2664047; Email: saksoy@cs.bilkent.edu.tr.

Ö. Çavuş is with Bilkent University, Department of Computer Engineering, Ankara, 06800, Turkey. Email: cavus@cs.bilkent.edu.tr.

textual and visual queries is described in Section III. Relevance feedback for iterative retrievals is presented in Section IV. Experimental results are discussed in Section V.

II. Image Representation

In this study, 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 [1].

III. Image Retrieval

Unfortunately, even when segments or grid cells are used, none of the existing feature extraction algorithms can always map visually similar regions to nearby locations in the feature space so images that are quite irrelevant to the query image often appear in the retrieval results. This problem, also referred to as the "semantic gap", has made interactive retrieval an important research problem to capture the high-level concept of similarity and subjectivity in human perception.

In our retrieval scenario, the user starts a retrieval 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 of videos extracted using automatic speech recognition techniques. These transcripts are obtained after processing the raw text using stemming, tagging, and frequency-based filtering steps. Closed caption text for videos can also be used for text-based querying if it is available. Alternatively, a random browsing of the shots can be used to generate the initial retrieval set.

Next, the user selects one or more of these shots as the visual query, and a content-based search is performed on the keyframes representing all shots in the database. The initial similarity between images is measured by appending all feature vectors for all grid cells and computing Euclidean distances between the resulting vectors. The initial iteration of visual querying sessions assumes that each feature in each grid cell has equal contribution to the computation of similarity. Upon being presented the results of this search, the user labels some of these shots as relevant (positive) and irrelevant (negative).

IV. Relevance Feedback

The feedback information is incorporated into the database 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

$$d = \sum_i \sum_j \sum_k w_{ijk} d_{ijk}. \quad (1)$$

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.

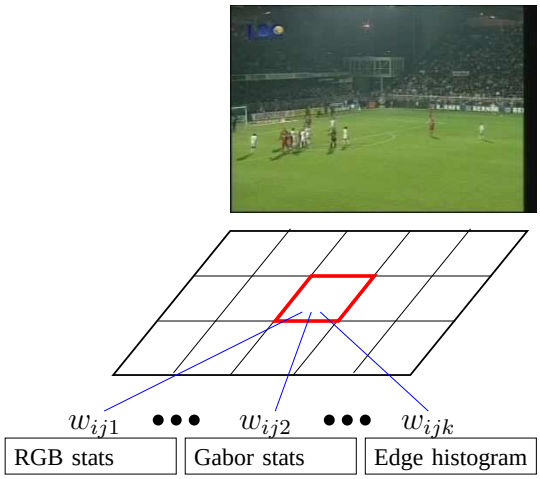


Fig. 1. A 3×5 grid layout and the corresponding weights for a particular grid cell. A 5×7 layout and 5 feature vectors are used in the experiments.

The weights are assigned uniformly in the first iteration. In earlier work [2], 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.

V. Experiments

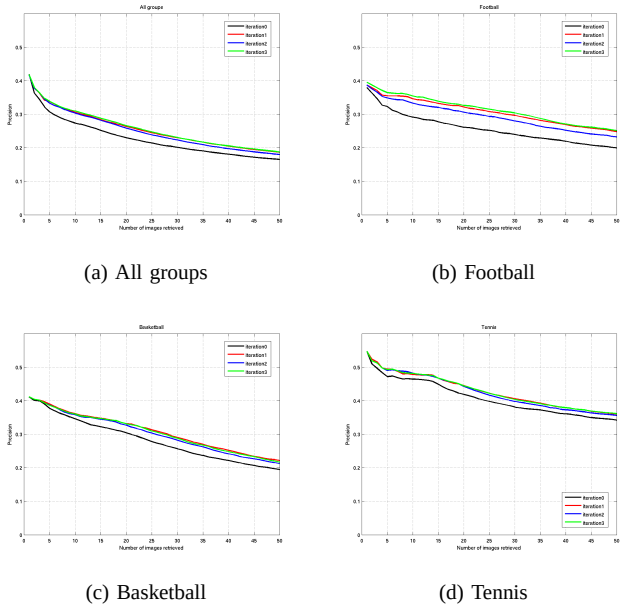
We evaluated the proposed techniques using 137 news videos from the development set of the TRECVID 2005 video retrieval evaluation [7]. As part of the common annotation effort, shots that were manually labeled by several TRECVID participants as containing sports such as football, basketball, golf, tennis, and American football were used as the ground truth among 43,907 shots in the TRECVID corpus. Example shots from each category are shown in Figure 2.

We have used this ground truth to automatically generate queries, sort approximately 44,000 shots for each query, and provide feedback to the system using the top 30 shots by automatically labeling each shot that belonged to the same ground truth group as the query as relevant and the remaining shots as irrelevant at each iteration. This process was repeated for each shot in the ground truth for 3 feedback iterations.

Figure 3 shows precision plots for different ground truth groups when different feature combinations were used. (Due to space limitations, only these plots can be shown.) Among all the combinations, using color features (RGB, HSV, LUV) and Gabor features gave the highest average precision. Adding edge orientation features did not improve the performance because there was no significant edge structure in the categories of interest. We expect that these features will be useful for other categories such as cities and offices.

The first iteration gave the largest increase in precision. Following iterations fluctuated around this first one. It is worth noting that feedback always improved the results with respect to the case without feedback for any combination of features.

Figures 4 and 6 show example retrievals for football and basketball queries, respectively, using color features. The corresponding weights for each feature in each grid cell are shown in Figures 5 and 7. These weights show that lower parts of the image (green field) were significant for color features for the football query whereas lower-right (score display) and top parts were more significant for the basketball query. The weights could successfully capture relative contributions of different features and spatial locations in computation of similarity using relevance feedback.



(a) All groups (b) Football

(c) Basketball (d) Tennis

Fig. 3. Precision vs. number of images retrieved for different ground truth groups and the overall performance for the original query (iteration 0) and the following 3 iterations when RGB, HSV, and LUV color features and Gabor texture features were combined.

References

[1] S. Aksoy and R. M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, May 2001.

[2] 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.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, August 2002.

[4] G.-D. Guo, A. K. Jain, W.-Y. Ma, and H.-J. Zhang. Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Transactions on Neural Networks*, 13(4):811–820, July 2002.

[5] T. S. Huang and Z. S. Zhou. Image retrieval with relevance feedback: from heuristic weight adjustment to optimal learning methods. In *Proceedings of IEEE International Conference on Image Processing*, volume 3, pages 2–5, Thessaloniki, Greece, October 2001.

[6] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. An efficient and effective region-based image retrieval framework. *IEEE Transactions on Image Processing*, 13(5):699–709, May 2004.

[7] U.S. National Institute of Standards and Technology. TREC video retrieval evaluation (TRECVID), 2005. http://www-nlpir.nist.gov/projects/trecvid/.

[8] Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 236–243, Hilton Head Island, South Carolina, June 2000.

[9] Y. Rui, T. S. Huang, and S-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, March 1999.

[10] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[13] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8:536–544, 2003.

Fig. 2. Example shots for different ground truth groups. The rows correspond to football, basketball, tennis, golf, and American football, respectively.

(a) Original query (19 correct)

(b) First feedback iteration (all 30 correct)

Fig. 4. Top 30 shots from a football query for one feedback iteration.

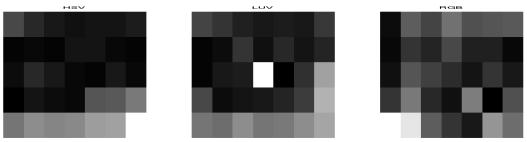


Fig. 5. Weights for different features (left to right: HSV, LUV, RGB) and grid cells for the football query. Brighter colors represent higher weights.

(a) Original query (14 correct)

(b) First feedback iteration (20 correct)

Fig. 6. Top 30 shots from a basketball query for one feedback iteration.

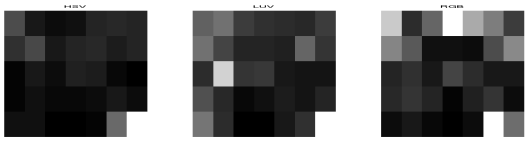


Fig. 7. Weights for different features (left to right: HSV, LUV, RGB) and grid cells for the basketball query. Brighter colors represent higher weights.