Interactive Object-based Retrieval Using Relevance Feedback

Sorin Sav, Hyowon Lee, Noel O'Connor, Alan F. Smeaton

Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland.
sorinsav@eeng.dcu.ie

Abstract. In this paper we present an interactive, object-based video retrieval system which features a novel query formulation method that is used to iteratively refine an underlying model of the search object. As the user continues query composition and browsing of retrieval results, the system's object modeling process, based on Gaussian probability distributions, becomes incrementally more accurate, leading to better search results. To make the interactive process understandable and easy to use, a custom user-interface has been designed and implemented that allows the user to interact with segmented objects in formulating a query, in browsing a search result, and in re-formulating a query by selecting an object in the search result.

1 Introduction

Automatic segmentation and indexing of objects such as persons, cars or buildings, represents one of the most active research areas in content-based image and video retrieval [1]. However, considering the interest in the problem and the variation of approaches and effort currently undertaken in this direction [2] [3] , progress is slow and performance accurate enough to be used in real applications still seems to be a distant goal. In the task of automatically segmenting and indexing objects in image/video content, the main difficulty is the diverse manifestations of an object in the image/video regardless of the object's inherent visual features such as colour, shape and texture. Factors such as different lighting conditions and camera angles and occlusions make the actual segmentation of an object extremely difficult, even before it can be accurately labeled. Considering this problem, one workaround solution we have been exploring is to use relevance feedback to take a human user's judgements on object definitions into account in retrieving objects. There is a long history of experimentation and successful use of relevance feedback in text-based information retrieval. This has included short-term modelling of a user's information need by dynamically updating the user's query formulation in mid-search as well as long-term modelling of user's needs by profiling his/her interests over time leading to personalisation. This has also been successfully applied to content-based retrieval [4] [5]

In this paper, we present an interactive, object-based search system that uses a novel query formulation mechanism and makes use of the user's query formulations as automatic feedback to the system in order to develop and refine the modeling of segmented objects in the database. As query formulation is the key element for getting feedback from the user in our approach, the system we have built incorporates a user interaction strategy at the front-end in which a user can efficiently and easily interact with segmented objects in video keyframes. The approach allows the user to highlight any segmented objects, select them, and then to use them for subsequent query formulation. The novelty of this work lies in using query formulations from users as implicit relevance feedback in developing more accurate object classes, the use of object matching in retrieval and the fact that we have built a system for users to interact with.

The remainder of the paper is organised as follows: in Section 2 we give an overview of the system explaining how objects are segmented and stored in the database and how user query formulation is used in an interactive session to refine object modelling in the database for subsequent retrieval. Section 3 describes the video object retrieval mechanism. Experimental results are presented in Section 4. Section 5 concludes the paper and outlines our plans for extending the system's capability and further refining the user-interface.

2 System Overview

Our system processes one object from each keyframe taken from each shot in the video and stores these in the database to be used in the retrieval process during an interactive search session (see Figure 1). We use keyframes automatically extracted from the TRECVid 2003 [6] test corpus, as well as images from the well known Corel test corpus.

For each keyframe, a semi-automatic object segmentation process was used to accurately segment one main object in the image. The segmentation tool used was previously reported in [7]. It allows fast and accurate automatic segmentation based on a small amount of user interaction that is easy to perform. The output can be iteratively refined in order to obtain very accurate object segmentations.

Once segmented, each object is automatically indexed by colour, shape and texture using the following well known MPEG-7 descriptors [8]: dominant colour descriptor, the compactness moment of the shape and the texture browsing descriptor. The motion feature as depicted in Figure 1 is not currently incorporated into the working system but the user-interface we have designed includes all four features for smoother upgrading of the underlying system in the near future. This completes the offline object segmentation and indexing process. Determining similarity among objects for retrieval purposes is done during interactive search without pre-computation as the system progressively receives more information from the user.

Query formulation is the core user interaction required to achieve more accurate search through iterative refinement of object modeling. Relevance feedback

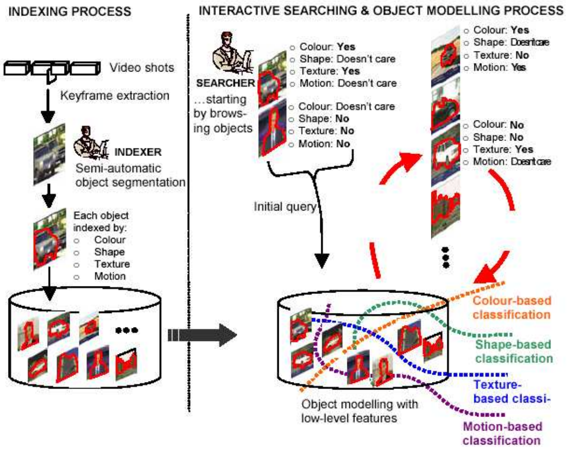


Fig. 1. System Overview showing off-line indexing and interactive searching

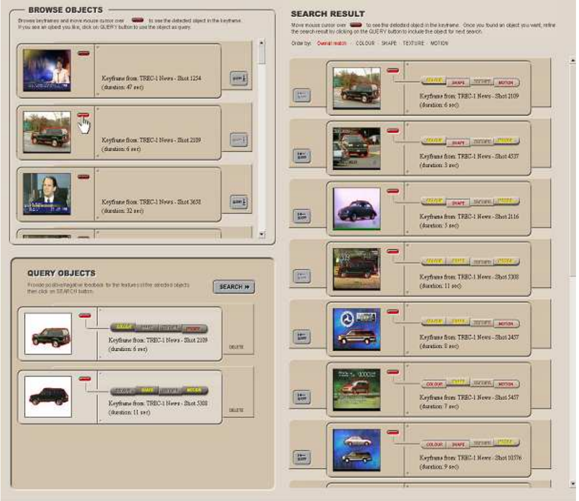
occurs each time a user formulates a query to search for objects. Figure 2 shows a screen from our interface after two iterations of query formulation and viewing of search results.

In Figure 2, after selecting an object, the user can then specify which low-level features (colour, shape or texture) of the specified object s/he is interested in. Each of the feature buttons toggles between positive, negative or neutral preferences for each feature.

3 Retrieval using User Query Formulation as Relevance Feedback

The initial query composed by the user is analysed in terms of the three low-level features (colour, shape and texture) and the degree of similarity between the query object's features and other objects in the database is computed. During interactive search, as more and more query formulation is conducted, the set of objects making up the input query becomes quite complex as it contains many objects, each of which represent positive or negative indications of the three features of the object the user is searching for. We assume the positive samples to be modeled by a mixture of Gaussian probability distribution functions (PDF) at feature level. Accordingly, each feature distribution is independently modeled as a Gaussian mixture, an assumption which is commonly used for image retrieval [9]. The feature vectors are modeled as a mixture of Gaussian distributions of the form:

$$f(X_i|\Phi) = \sum_{j=1}^{k} \pi_j f_j(X_i|\theta_j) \quad . \tag{1}$$



BROWSE OBJECTS

Browse keyframes and move mouse cursor over to see the detected object in the keyframe. If you see an object you like, click on QUERY button to use the object as query.

Keyframe from: TREC-1 News - Shot 1254 (duration: 47 sec)

Keyframe from: TREC-1 News - Shot 2109 (duration: 6 sec)

Keyframe from: TREC-1 News - Shot 3658 (duration: 32 sec)

QUERY OBJECTS

Provide positive/negative feedback for the features of the selected objects then click on SEARCH button.

Keyframe from: TREC-1 News - Shot 2109 (duration: 6 sec)

Keyframe from: TREC-1 News - Shot 5308 (duration: 11 sec)

SEARCH RESULT

Move mouse cursor over to see the detected object in the keyframe. Once you found an object you want, refine the search result by clicking on the QUERY button to include the object for next search.

Order by: Overall match · COLOUR · SHAPE · TEXTURE · MOTION

Keyframe from: TREC-1 News - Shot 2109 (duration: 6 sec)

Keyframe from: TREC-1 News - Shot 4537 (duration: 3 sec)

Keyframe from: TREC-1 News - Shot 2116 (duration: 5 sec)

Keyframe from: TREC-1 News - Shot 5308 (duration: 11 sec)

Keyframe from: TREC-1 News - Shot 2457 (duration: 8 sec)

Keyframe from: TREC-1 News - Shot 5457 (duration: 7 sec)

Keyframe from: TREC-1 News - Shot 10576 (duration: 9 sec)

Fig. 2. User interface after two iterations of query composition using objects

where:

$$f(X_i|\theta_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\sum_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_i-\mu_j)^T \sum_j^{-1}(X_i-\mu_j)} \quad . \tag{2}$$

is the probability density function for cluster $j, \theta_j = (\mu_j, \sum_j)$ is the set of parameters for density function $f_j(X_i|\theta_j)$, μ_j is the mean of cluster j, π_j is the mixing proportion of cluster j subject to the condition $\pi_j \geq 0$ and $\sum_{j-1}^{k} \pi_j = 1$ where k is the number of components. X_i is the vector for either colour, shape or texture, $\Phi = (\pi_1, \pi_2 ... \pi_k, \theta_1, \theta_1 ... \theta_k)$ is the set of all parameters. Here $f(X_i|\Phi)$ is the probability density function given the colour, shape or texture, of the labelled object X_i for each of the three features of a query object labelled by the user.

As the number of components in the mixture becomes larger, the model tends to follow the real distribution of the positive samples' features more accurately. However, maintaining and operating using a large model is increasingly difficult and therefore there is a need to restrain the model size. The model is built on the sample objects indicated by the user and unlabeled data. A minimum description length (MDL) constraint is used to ensure that the Gaussian mixture has the minimum number of components that correctly classifies the labeled (user indicated) set of objects without including a significant number of negative samples (model outliers). The number of components in the mixture is increased, when the user indicates new samples, only if the following expression is true:

$$\alpha[\log f(X|\Phi)^{(t+1)} - \log f(X|\Phi)^{(t)}] > \beta(N^{(t)} - N^{(t+1)}) \quad . \tag{3}$$

where t is the number of Gaussian components in the mixture, $logf(X|\Phi)$ is the log-likelihood function and N is the number of negative samples (outliers) contained within the modelled PDF. The α and β parameters are system weighting factors currently set to: $\alpha = 0.23$, $\beta = 0.07$.

The estimation-maximisation (EM) algorithm [10] is employed to estimate the PDF in the feature space in connection with the MDL constraint given by (3). The maximization is performed by the following iteration:

$$E[z_{ij}] = p(z_{ij} = 1|X, \Phi_{(t)}) = \frac{\pi_j^{(t)} p_j(X_i|\Phi_j^{(t)})}{\sum_{s=1}^{k} p_s(X_i|\Phi_s^{(t)})\pi_s^{(t)}} \ . \tag{4}$$

$$\pi_j^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} E[z_{ij}], \mu_j^{(t+1)} = \frac{1}{N\pi_j^{(t+1)}}\sum_{i=1}^{N} E[z_{ij}]X_i \ . \tag{5}$$

$$\Sigma_j^{(t+1)} = \frac{1}{N\pi_j^{(t+1)}}\sum_{i=1}^{N} E[z_{ij}](X_i - \mu_j^{(t+1)})(X_i - \mu_j^{(t+1)})^T \ . \tag{6}$$

where $E[z_{ij}]$ is the expected value of the probability that the data belongs to cluster j, and $\sum_{i=1}^{N} E[z_{ij}$ is the estimated number of data points in class j. At each iteration, the model parameters are re-estimated to maximize the model log-likelihood, $f(X|\Phi)$, until convergence.

At each retrieval iteration, the Mahalanobis distance [11] from each Gaussian feature cluster to the existing objects in the database is computed as a measure of similarity (a minimum distance classifier) and the objects in the database are presented to the user as a ranked list in descending order of the cumulative similarity score $S(x)$ where each feature is weighted in direct proportion to the number of its positive samples indicated by the user. The Mahalanobis distance is expressed as:

$$r^2 = (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) \ . \tag{7}$$

where x is the vector for either colour, shape or texture, μ_i is the mean vector, and Σ_i^{-1} is the diagonal covariance matrix for each of the colour and shape clusters. The weighting scheme favors the feature more often indicated as positive because its repeated occurrence suggests a larger incidence of similar objects in the database. The cumulative similarity score is expressed as:

$$S(x) = \lambda_{colour}S_{colour}(x) + \lambda_{shape}S_{shape}(x) + \lambda_{texture}S_{texture}(x) \ . \tag{8}$$

where $S_{feature}$ is the Mahalanobis distance for the given feature, and λ is computed as:

$$\lambda_k = \frac{P_k}{P_k + P_l + P_m} \ . \tag{9}$$

with k, l, m being the features colour, shape and texture.

4 Experimental results

In order to evaluate the performance of the system we designed a retrieval experiment using 12 classes of objects, each class containing 50 objects. The objects classes used are: balloon, boat, butterfly, car, eagle, flower, horse, motorcycle, people, plane, shark, tiger.

Experiments were performed with an expert user selecting an initial query object and providing negative/positive feedback. For each query iteration a positive example was added in the query formulation, a negative example was added every second iteration. The query session for each object class was conducted for 5 iterations, therefore for each object class 5 positive examples and 2 negative examples were provided over 5 iterations. The mean precision-recall curves obtained are shown in Figure 3. Since representing 12 curves on the same graph becomes confusing, we present the precision-recall curves grouped on four sub-images for every three classes taken in alphabetical order. In order to provide a easy comparison between object classes, each sub-image contains the mean precision versus recall curve computed by averaging the results over the entire 12 classes.

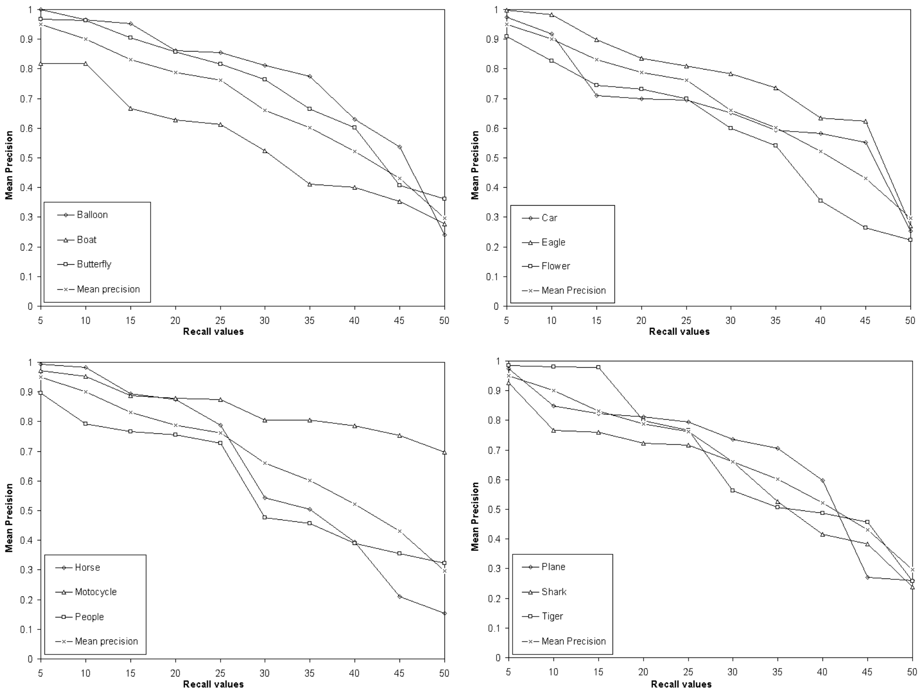


Fig. 3. Mean precision vs recall curves for 12 object classes

The precision-recall curves show a relatively slow decay with increasing recall. The optimal values seem to be located around values of recall of 15-25 images/objects out of 50 objects per class, which seems to prove the effectiveness of the presented system. However, it is premature to generalise before performing comparisons against other retrieval systems on a common test set and with multiple users in the retrieval loop.

5 Conclusions and future work

In this paper we introduced an object-based video search system that features interactive query formulation using colour, shape and texture of an object. Iterative query/browsing incrementally improves object modelling in the data-base. The actual segmentation of objects from keyframes was supervised in order to provide accurate object sets and to better illustrate our retrieval approach in which the matching among objects (i.e. relating all similar objects in the database) can be helped using the user's query formulation history as feedback.

In its present form our system may not to be suitable for a realistic context, but the point of developing it was to demonstrate how an object-based query formulation mechanism could be realised to help dynamically refine the object model in the database and enhance retrieval.

We are working on several improvements including making object segmentation from each keyframe fully-automatic. Segmenting more than one object from each keyframe is also targeted for future work. Currently, a keyframe from a shot is used to segment objects: however, a more complete solution would be to use all frames within the shot, which could further provide additional information on the object based on its movement and trajectory.

6 Acknowledgments

The support of the Enterprise Ireland Informatics Initiative is gratefully acknowledged. Part of this work was supported by Science Foundation Ireland under grant 03/IN.3/I361

References

1. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R. Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1349-1380, 2000.
2. Smith, J.R., and Chang, S.F,. VisualSEEK: a fully automated content-based image query system. ACM Multimedia, Boston, November, 1996.
3. Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., Malik, J. Blobworld: A System for Region-Based Image Indexing and Retrieval. Proceedings of the Third International Conference on Visual Information and Information Systems, 1999.
4. Rui, Y., Huang, T. S. and Mehrotra S. Content-based image retrieval with relevance feed-back in Mars. Proceedings of IEEE International Conference on Image Processing ICIP, 1997.

5. Yan, R., Hauptmann, A. and Jin, R. Multimedia Search with Pseudo -Relevance Feedback. Proceedings of International Conference on Image and Video Retrieval CIVR 2003, Urbana, IL, July 24-25, 2003.
6. http://www-nlpir.nist.gov/projects/trecvid
7. O'Connor, N., Adamek, T., Sav, S., Murphy, N. and Marlow, S. QIMERA: A Software Platform for Video Object Segmentation and Tracking. Proceedings of the 4th Workshop on Image Analysis for Multimedia Interactive Service (WIAMIS 2003), London, U.K., April 9-11, 2003.
8. Salambier, P. and Smith, J.R. MPEG-7 Multimedia Descriptions Schemes. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, pp. 748-759, June 2001.
9. Qian, F., Li, M., Zhang, L., Zhang, H.J. and Zhang, B. Gaussian mixture model for relevance feedback in image retrieval. Proceeding of IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August, 2002.
10. Moon, T.K. The Expectation-Maximisation Algorithm. IEEE Signal Processing Magazine, November, 1996.
11. Fessant, F., Aknin, P., Oukhellou, L., and Midenet, S. Comparison of supervised self-organizing maps using Euclidian or Mahalanobis distance in classification context. Proceedings of the 6th International Work Conference on Artificial and Natural Neural Networks (IWANN2001), Granada, Spain, June 13-15, 2001.