ASSOCIATING LOW-LEVEL FEATURES WITH SEMANTIC CONCEPTS USING VIDEO OBJECTS AND RELEVANCE FEEDBACK

Sorin Sav, Noel O'Connor, Alan Smeaton, Noel Murphy

Centre for Digital Video Processing, Dublin City University, Ireland

ABSTRACT

The holy grail of multimedia indexing and retrieval is developing algorithms capable of imitating human abilities in distinguishing and recognising semantic concepts within the content, so that retrieval can be based on ”real world” concepts that come naturally to users. In this paper, we discuss an approach to using segmented video objects as the mid-level connection between low-level features and semantic concept description. In this paper, we consider a video object as a particular instance of a semantic concept and we model the semantic concept as an average representation of its instances. A system supporting object-based search through a test corpus is presented that allows matching pre-segmented objects based on automatically extracted low-level features. In the system, relevance feedback is employed to drive the learning of the semantic model during a regular search process.

1. INTRODUCTION

The continuous expansion of multimedia archives has resulted in an increasing demand for effective information management systems. Content-based information retrieval (CBIR) systems are required to assist humans in locating and retrieving relevant content. Ideally, retrieval should be performed using concepts and queries that come natural to humans. However, retrieval systems often operate with low-level representations [1], more accessible for machine processing, whilst humans perceive multimedia at the semantic level [2]. Similarity measures constructed on low-level features [3] [4] [5], such as colour, texture and shape, will only capture analogous features and to move beyond this, concept identification is required. However, unconstrained object recognition and subsequent concept identification is considered to remain far beyond the capabilities of the research community in the near future [6]. To make matters worse, mapping between feature level descriptions and human concepts is ambiguous due to multiple concepts sharing similar distributions of features.

Relevance feedback has been proposed as an approach that allows the user to implicitly construct a feature-concept mapping by directly influencing the retrieval process. Originally in image retrieval, the knowledge obtained from user interaction was regarded as specific to each query and therefore discarded at the end of the session. Recent approaches, however, attempt to further use this information by inferring the semantic space [7] driving the user's actions. Furthermore, the inferred semantic space, could be later used in building connections among the images in the archive [8] or for propagating annotation terms [9] in order to subsequently aid retrieval for other users.

Humans can effortlessly assess visual content in mid-level terms (colour, shape, texture), but find it difficult to detect and express variations in low-level features (e.g. colour moments, shape moments, texture descriptors). In retrieval systems that work only on low-level features, feedback from users is generally limited to the entire content of the image. In systems that divide the image into regions (automatically or by user interaction) feedback can be given for each separate region in a partition. It has been argued that feedback effectiveness increases as the partitions employed more closely express the semantics of the scene [10]. Therefore, in this paper we consider extracting concept descriptions based on low-level features directly related to human perception of the objects present in an image.

The remainder of this paper is organized as follows. Section 2 provides an overview of the overall approach. Object feature representation is described in Section 3. The use of relevance feedback is explained in Section 4 and experimental results are presented in Section 5. We present conclusions and ideas for future work in Section 6.

2. ASSOCIATING OBJECTS WITH SEMANTIC CONCEPTS

We approach retrieval from a combined perspective: as short term learning for immediate increase of retrieval performance and as long term learning with the view of modelling the semantics of the query. Short term improvement of the retrieval performance means presenting more relevant items to the user, while for long term learning it is critical to obtain a large set of positive and negative training examples.

The proposed system uses the semantic video objects extracted from the keyframe associated to the video shots in the archive. When new shots are added to the archive for each keyframe the most representative video objects are extracted using the interactive tool for segmentation described in [11]. Each object is automatically described in terms of colour and shape (see Section 3). Colour and shape are used as they are independent features directly related to human perception of objects. For images with no obvious video objects, the colour distribution of the entire image is used, as typically such images represent scenery with no foreground objects. In this case, shape information is not considered.

During the retrieval process, clusters of features are associated to relevant/irrelevant objects based on user interaction. The relevance feedback approach employed (see Section 4) creates clusters of low-level features. The clusters that cover relevant objects are associated with the semantic concept conveyed by the keyword used to identify the query.

3. OBJECT FEATURE REPRESENTATION

3.1. Color

To represent colour we adopted the MPEG-7 dominant colour descriptor (DCD) [1] as used in many retrieval systems. In order to adapt this feature to objects as oppose to images only the interior of the object is considered.

The recommended distance to be used with DCD is [12]:

$$D_{DCD}(Q,I) = \left(\sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2\alpha_{1i,2j} \right)^{1/2} \tag{1}$$

where N is a set of colour vectors c_i, and p_i their percentages. The similarity coefficient $\alpha_{k,l}$ between two RGB color vectors c_k and c_l is calculated as:

$$\alpha_{k,l} = \begin{cases} 1 - \frac{D_{k,l}}{D_{max}}, & D_{k,l} \leq T_d \\ 0, & D_{k,l} > T_d \end{cases} \tag{2}$$

where $D_{k,l} = \parallel c_k - c_l \parallel$ represents the Euclidian distance between two colour vectors, $T_d = 20$, $\alpha = 1$, and $D_{max} = \alpha T_d = 20$ as suggested in [13].

3.2. Shape

Shape description and similarity is an extremely complex research topic – 2D projection on the image plane, elastic deformations of the object, and diversity of shapes in which instances of the same semantic object appear in the real world are common problems that must be considered for shape similarity. In this initial work, we use a relatively simple shape descriptor corresponding to the compactness moment γ [14]:

$$\gamma = \frac{P_2}{4\pi A} \tag{3}$$

where A is the area and P perimeter of the video object. This is a simple and robust descriptor that can indicate a degree of shape similarity.

4. BUILDING FEATURE CLUSTERS

Since the object descriptors are independent, the probability distribution of a semantic concept over an image descriptor space can be modeled as a Gaussian mixture model:

$$p(\varepsilon) = \sum_{j=1}^{N} \alpha_j p_j(\varepsilon|j) \tag{4}$$

where $\sum_{j=1}^{N} \alpha_j = 1$. Each mixture component is a Gaussian with mean μ and covariance matrix Σ:

$$p(\varepsilon|j) = \frac{1}{2\pi|\Sigma_j|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\varepsilon-\mu_j)^T \Sigma_j^{-1} (\varepsilon-\mu_j)} \tag{5}$$

At the start of the retrieval session the user selects a query image, and thus implicitly the extracted object associated with that image. The colour and shape of the query object are considered as the means μ of two independent Gaussian clusters, one in the colour space, the other one in the shape space. An arbitrary variance Σ is assumed for the initial clusters. The mean μ and variance Σ are then subsequently estimated from user labelled examples. Images containing objects similar to the query object are retrieved using the Mahalanobis distance and presented to the user explicitly rated as positive or negative instances relative to the query. We argue that presenting rated results makes the user more attentive to correcting by feedback erroneousness results. Thus, within the list of relevant results we deliberately insert low confidence items in order to elicit user feedback.

In conformity with the Minimum Description Length (MDL) principle we assume the best semantic model to be the Gaussian mixture of low-level features with the minimum number of components that correctly classifies the labeled set of images. Intuitively, this occurs when the majority of items in a cluster are similarly labelled (positive/nega–tive) and the cluster contains no oppositely labelled items. Clusters of positively labelled items are assumed to generate positive instances whilst clusters of negative items are assumed to generate negative instances. This is a weak assumption, however, as labelled items are sparsely distributed within the feature space. Therefore, we seek to verify the validity of the assumption by inserting negative samples within the cluster as relevant results (low-confidence or negative instances), expecting the user to label them appropriately. The Gaussian mixture is updated, when feedback is

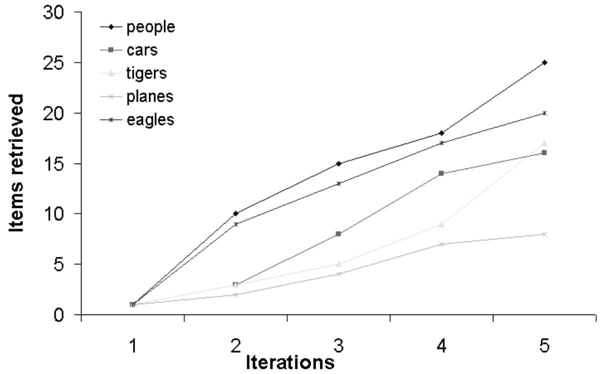

Fig. 1. Retrieved relevant items against iterations for each predefined category of object

available, based on a estimation-maximisation (EM) technique [15].

The EM algorithm consists of two steps an expectation step (E) and a maximisation step(M). In the E step a given number of clusters are build based on current estimate of the positive items distribution in the feature space. In the M step the underlying distributions of the positive items are re-estimated. These two steps are iterated until convergence.

5. EXPERIMENTAL RESULTS

For experimental purposes, we created a test corpus of objects culled from 2000 images from the Corel dataset and 1137 keyframes extracted from 2 hours of TV broadcast drawn from the TREC 2003 [16] corpus. In each image a single dominant object was manually segmented. We defined five categories of objects – people, cars, tigers, planes and eagles – and manually classified all segmented objects to one category in order to create a ground truth. Low-level features were extracted for all objects. Experiments were performed with an expert user selecting an initial query object and providing negative/positive feedback. Each query session was stopped after 25 feedback iterations.

Figure 1 shows the number of relevant items (as judged against the ground truth) retrieved against the number of iterations for each category. As can be seen the retrieval performance is better in the *people* category due to the similarities of object shape. In the *eagles* category there is large variation in shape and consequently the retrieval performance is lower.

Figures 2, 3 and 4 show the first 16 images retrieved for the *people, tigers* and *eagles* categories with the segmented object indicated in each case. In each case, the image in the upper left corner is the initial query image.

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented an approach to automatically associating low-level features extracted from segmented video objects to semantic concepts. The features used are the colour and shape of the objects. We use video objects as the mid-level link between low-level features and semantic concepts whereby we consider a video object as a particular instance of a semantic concept. Relevance feedback is employed to model the semantic concept as an average representation of its instances

Our main efforts in the future will focus on using more features other than those reported here and expanding the range of experiments and the size of the image corpus. Other work will concentrate on determining a criterion for selecting the minimum number of feature clusters that satisfactorily cover the instances of a given concept.

7. ACKNOWLEDGMENTS

The support of the Enterprise Ireland Informatics Initiative is gratefully acknowledged. This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

8. REFERENCES

[1] P. Salambier and J.R. Smith, “MPEG-7 multimedia descriptions schemes,” *IEEE Transactions Circuits and Systems for Video Technology*, vol. 11, pp. 748–759, June 2001.

[2] J.R. Smith and S.F. Chang, “Visualseek: A fully automated content-based query system,” *Proccedings of Fourth International Conference on Multimedia, ACM*, pp. 87–92, 1996.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, “Query by image and video content: The QBIC system,” *IEEE Computer*, , no. 28, pp. 23–32, September 1995.

[4] W.Y. Ma and B.S. Manjunath, “Netra: A toolbox for navigating large image databases,” *ACM Multimedia Systems*, , no. 7, pp. 184–198, 1999.

[5] Y. Rui, T.S. Huang, S. Methrotra and M. Ortega, “A relevance feedback architecture for content-based multimedia information retrieval system,” *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 82–89, 1997.

[6] S. Santini and R. Jain, “Visual navigation in perceptual databases,” *Proccedings of International Conference on Visual Information Systems, San Diego, CA*, December 1997.

[7] X. He, O. King, W.Y. Ma, M. Li and H.J. Zhang, “Learning a semantic space from user’s relevance feedback for image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 39–48, January 2003.

[8] M. Lee, W.Y. Ma and H.J. Zhang, “Information embbeding based on user’s relevance feedback for image retrieval,” *Proceedings of SPIE Multimedia Storage and Archiving Systems IV, Boston*, September 1999.

[9] B. Li, K. Goh and E.Y. Chang, “Confidence-based dynamic ensamble for image annotation and semantic discovery,” *Proceeding of the 11th ACM International Conference on Multimedia, Berkeley, CA, USA*, November 2003.

[10] S.F. Chang, W. Chen and H. Sundaram, “Semantic visual templates: Linking visual features to semantics,” *Proceeding of IEEE International Conference on Image Processing 1998, Chicago, IL, USA*, October 1998.

[11] N. O’Connor, T. Adamek, S. Sav, N. Murphy and S. Marlow, “QIMERA: A software platform for video object segmentation and tracking,” *Proceeding of the 4th European Workshop on Image Analysis for Multimedia Interactive Services, London, UK*, April 2003.

[12] B. Manjunath, P. Salambier and T. Sikora, *Introduction to MPEG-7: Multimedia content description interface*, Wiley, New York, USA, 2002.

[13] A. Kushki, P. Androutsos, K.N. Plataniotis and A.N. Venetsanopoulos, “Query feedback for interactive image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 644–655, May 2004.

[14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.

[15] T. K. Moon, “The Expectation-Maximisation Algorithm,” *IEEE Signal Processing Magazine*, pp. 47–60, November 1996.

[16] TrecVid 2003 web site:
http://www nlpir.nist.gov/projects/tv2003/tv2003.html.

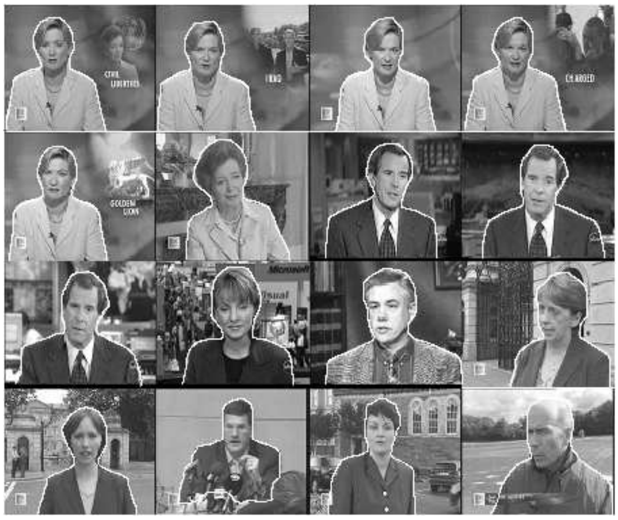

Fig. 2. Items retrieved in the *people* category

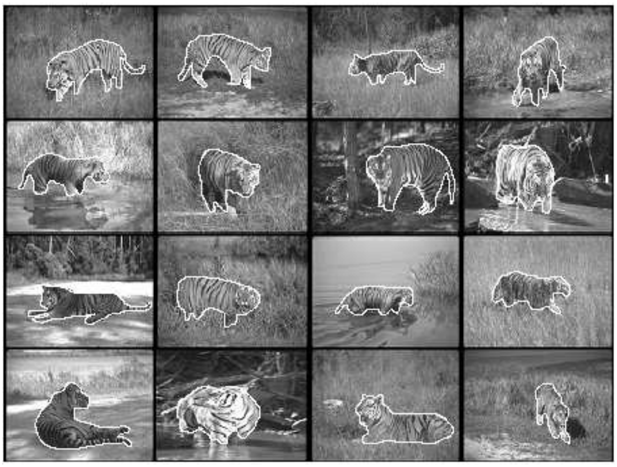

Fig. 3. Items retrieved in the *tigers* category

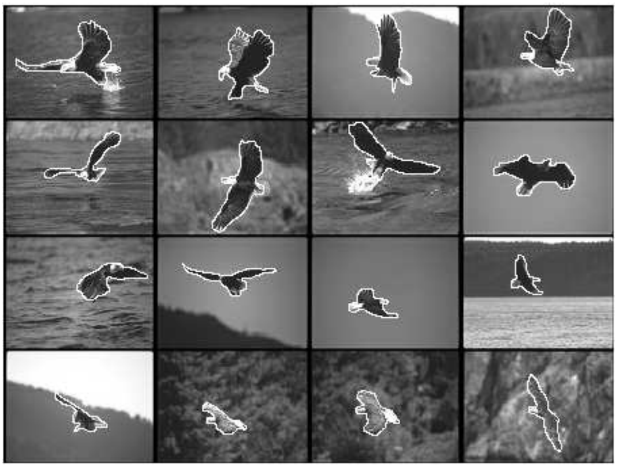

Fig. 4. Items retrieved in the *eagles* category