REGION-BASED RELEVANCE FEEDBACK IN CONCEPT-BASED IMAGE RETRIEVAL

José Torres[1,3], Alan Parkes[1], Luís Corte-Real[2]

[1]Lancaster University, {torres, app}@comp.lancs.ac.uk
[2]Faculdade de Engenharia da Universidade do Porto/Inesc-Porto, lreal@inescporto.pt
[3]Universidade Fernando Pessoa/Inesc-Porto

ABSTRACT

The image retrieval approach described in this paper combines two layers, the conceptual and the feature-based. It works at the region or visual object level and uses a thesaurus to explore associations between text and image content and enable concept-level queries. The solution described uses relevance feedback both at the query session level and to discover new associations between text and image.

Particular emphasis is given in the paper to the limitation of the typical relevance feedback solutions presented and to the benefits of using the relevance feedback in conjunction with the two-layer model presented. Some preliminary results are reported using the prototype VOIR - Visual Object Information Retrieval - system.

1. INTRODUCTION

The continuing rapid growth in size of digital image databases leads to the increasing need for retrieval systems to satisfy more complex and sophisticated queries. The large scale of image databases means that manually annotating the images is infeasible. Content-based image retrieval systems attempt to solve this problem by automating the process of image indexing. Nevertheless, as demonstrated by several user studies [1], [2], users are also interested in searching the images at a conceptual level, not only in terms of colour, texture or shape.

A key requirement for developing future image retrieval systems is to explore the synergy between humans and computers [3]. Relevance feedback is a technique that engages the user and the retrieval system in a process of symbiosis. Following the formulation of the initial query, for subsequent iterations of query refinement, the system presents a set of results and the user evaluates the results in order to refine the set of images retrieved to his or her satisfaction.

This paper analyses the use of relevance feedback in image retrieval and presents a framework that takes into consideration a two-layer model: the *conceptual* and the *visual* layer.

The paper is organised as follows. Section 2 discusses related work in relevance feedback. In section 3, details of the idealized framework are presented in addition to the two-layer model of description of visual items. Section 4 describes the Visual Object Information Retrieval (VOIR) system and some experimental results obtained. The final section gives concluding remarks and a brief discussion of future work.

2. RELATED WORK

The involvement of the user in an iterative process of information retrieval is not new and was incorporated in the text retrieval system SMART [4]. The adoption of user relevance feedback for refinement of queries over image repositories has been a topic of active research in recent years.

The MARS image retrieval system [5] is one of the most cited in the literature. In the model of relevance feedback used in MARS, at each iteration the system tries to calculate a new ideal query point. This calculation is based on the user's evaluation of the results of the previous iteration. Two methods are used to implement this technique: query point movement and query re-weighting. The form (shape and orientation) of the hyper-surfaces used in the distance function was later improved to allow generalized hyper-ellipsoids in the *MindReader* system [6]. Rui et al. [7] present a more general model that includes previous models as special cases.

The systems in the previous paragraph embody the assumption that user expectations or target images are directly mapped onto the adopted feature space. As well as estimating the ideal parameters or weights for each axis of the hyper-ellipsoid, such systems also adopt a query point calculation method that attempts to compute the ideal single point in the feature space in order to retrieve the nearest images to it. As explained in the next section, this approach is limited, since in a semantic level query the user may want results associated with several visual representations.

The Falcon system for query by multiple examples [8] proposes one parametrical "aggregate dissimilarity" function that attempts to reduce the problem of using

single point queries, as discussed later, while taking into account the several distances between the candidate point x and the multiple good objects g_i. Experiments have supported the intuition that the best results are achieved when the function mimics a fuzzy OR.

The *iFind* system [9] features a scheme to associate user-entered keywords from an uncontrolled vocabulary with corresponding images. Each of these associations has a corresponding weight that is heuristically updated during subsequent use of the system. In parallel to this, the system uses a low-level feature based relevance feedback scheme based on the work described by [7].

One region-based relevance feedback system has yielded promising results from using region segmentation and representation [10].

In [11] a method that learns the relations between images based on the user feedback is presented. These relations are stored in one undirected graph that constitutes the "Semantic Layer". A further undirected graph that constitutes the "Visual Layer" is used to store pairs of images that have a (low-level) visual similarity above a certain threshold. The retrieval is performed using a process of link analysis of the graphs. The method described does not use keywords.

The method proposed in [12] also uses relevance feedback to split and merge image clusters that are formed in the low level feature space. Relations between the clusters are expressed using a correlation matrix. The existing clusters as well as the correlation matrix are updated during iterative use of the system.

3. PROPOSED APPROACH

The method we present embodies the assumption that the target images of the user are fundamentally associated with concepts. So, for example, the user may be interested in finding images of cars or chairs or airplanes, and not especially concerned with a particular colour or orientation. Given this assumption, the system may attempt to find the target images in several visual categories or visual representations associated with the desired concept or semantic category, following the model presented in section 3.1. Content-based features extracted from each visual item, such as colour, are also considered during the search process.

3.1 The Conceptual and the Visual Layer

Two levels of similarity are discussed in this section: *conceptual* and *visual*. Most image retrieval systems include a feature level where each visual item v_i has an associated automatically extracted feature f_i as its signature. This model is deficient in that, in themselves, the features are not usually sufficient to support the correct discrimination of *conceptual* similarity between distinct visual items.

The conceptual level typically implies designating the item as a member of a conceptual category. Suppose, for example, that we have in our database two images of cars, one representing a top view of a blue saloon car and other representing a front view of red sports car. Although the appearance of each is different, they both belong to the conceptual "car" category. On the other hand, two or more extremely similar visual items according to a given feature space and a distance function are considered visually similar, although they could belong to radically distinct semantic categories.

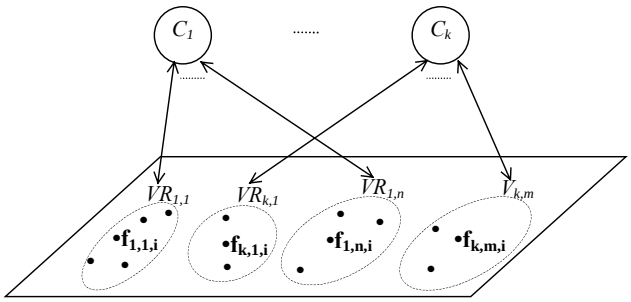

Figure 1 - layered representation of the hierarchical relation between conceptual categories and visual categories

Figure 1 shows a two-layer model separating conceptual categories at the upper layer from the visual categories or visual representations. Each conceptual category, C_i, is composed of several visual categories or visual representations, $VR_{i,j}$. Each visual representation is composed of several items that are represented by a point in the feature space (bi-dimensional in the figure). The ellipses delimit the visual items that belong to the same visual representation. It illustrates why it is not possible to join visual and conceptual categories in a single category. To try to use the feature space as a metric space to infer conceptual similarity between visual items using their signatures is simply impossible since the problem is ill defined in nature. As can be seen in the example, although the signature pair ($f_{1,1,i}$; $f_{k,1,i}$) belong to different conceptual categories, they are nearer than the signature pair ($f_{1,1,i}$; $f_{1,n,i}$) that belong to the same conceptual category.

In the adopted approach each concept C_i is associated with a corresponding label or textual term in a one-to-one mapping. The permissible textual terms (labels) that can be associated with the visual objects are extracted from a controlled vocabulary, in this case a textual thesaurus. The use of a textual thesaurus reduces inconsistency in term assignment and provides a knowledge structure that can be explored during the searching process.

One important feature of the idealized approach is the description of visual items at the region level. The solution proposed uses a region-based approach for

representation, query and retrieval of images. It is assumed that the images are already segmented into regions before being indexed. During the indexing operation, each region is associated with a feature vector representing low-level features such as colour, texture and shape. During query formulation, after the user chooses the textual term from the thesaurus that represents the desired concept, he or she can select one of the visual regions already previously associated with the term to be used as the example during the content-based query.

The relevance feedback information provided by the user supports refinement of the results, but can also be used to improve the behaviour of the image retrieval system in subsequent sessions. In the latter context, the system is said to be evolving over time. The proposed approach takes advantage of the relevance feedback information at both levels.

3.2 Relevance Feedback at the session level

In each query session, the system implements a relevance feedback mechanism that attempts to move the query point towards the good points and away from the bad points. It also attempts to reweigh the query so as to increase the weight of the more discriminating features. These two methods have been used elsewhere [5]. The novelty of our approach is that, instead of limiting the number of query points to just one, it expands the query by using additional query points in the feature space that are related with the same semantic category. This follows the model discussed in section 3.1, above.

When a new relevant example f_i is indicated by the user, a *Boolean* function will indicate if the designated point belongs to the same visual category of the evaluated visual item f_j or not. If this is the case, the new point will be considered as one more positive point of the evaluated item. If is not the case, this point will be considered as the seed of another visual category to be added to the current query.

The current implementation of the mentioned function, essentially compares the distance $D_{ji} = distance(f_j, f_i)$ with $D_{jk} = distance(f_j, f_k)$ where $f_k \in F_K$ the set of all visual items whose category C_k is different of the category C_i of point f_i. Basically the query expansion is done if $(D_{ji} / D_{jk}) > thr$, where *thr* is a pre-defined threshold level.

3.3 Concept learning using Relevance Feedback

As stated earlier, each visual item represents a region of an image. The association between textual terms and visual items is characterized by having a normalized degree of confidence *d_conf* where the attribute *d_conf* $\in$ [0, 100]. This association is of fundamental importance since it constitutes the outcome of the process of concept learning. It can be done manually or automatically. In the first case *d_conf* is set to its maximum value (100), in the second case it will be defined or updated algorithmically.

The critical evaluation of the image results by the user during query sessions is used to create or update the existing associations. The outcome of this is that the system gradually learns associations between visual regions and labels from the textual thesaurus. The more the system learns, the more accurate and faster are the subsequent query sessions.

In the implementation used to carry out the experiments, the visual categories, used in the concept learning process, were defined off-line using a clustering algorithm that took low-level features extracted from each region as its input data. The automatic updating of the associations between term and visual item is done periodically after the query sessions or following new manually added associations. The updating process affects all the visual items that belong to the same visual category as the visual item whose situation was changed either because was explicitly associated with a keyword or because was evaluated during a query iteration.

4. EXPERIMENTS

The experiments reported in this section were conducted using the VOIR image retrieval system [13] as the prototype platform. The current implementation of VOIR uses the Australian Pictorial Thesaurus [14] as the basis of the textual thesaurus.

The VOIR indexer module associates with each segmented region a collection of automatically extracted numerical properties. The low-level descriptors used in this experiment were colour histograms in the L*a*b* space, a texture descriptor adapted from the MPEG-7 Edge Histogram Descriptor [15], and the first five region central moments.

The currently adopted partition at the visual layer is static and was obtained as the outcome of the off-line clustering procedure during initialisation. The clustering algorithm adopted was the *Classit* [16].

The image collection used in the experiments was a database containing "ground-truth", human-authored image segmentations made available for research use [17]. The database is composed of 300 images from the Corel dataset all labelled according to diverse categories such as *animals*, *plants*, *people*, *landscape earth features* (mountains, bushes, etc.), *manufactured objects* (airplanes, etc.), and so on. The total number of image segments is around 3100 representing an average of approximately 10-11 regions per image. The number of different keywords used in the categorisation is around 300.

The average number of images per each keyword, for the first 20 most frequent keywords is around 25. From

this group, 4 were chosen randomly to conduct the experiments.

Figure 2 shows the average precision obtained regarding two distinct moments of learning, where *t0* represents the status when there is no text-image association and *t1* represents a more advanced state of knowledge after some level of usage of the system. The limitation observed in the precision value is due to the fact that the number of relevant images per keyword used is, on average, about half the number of the images returned by the system in each iteration.

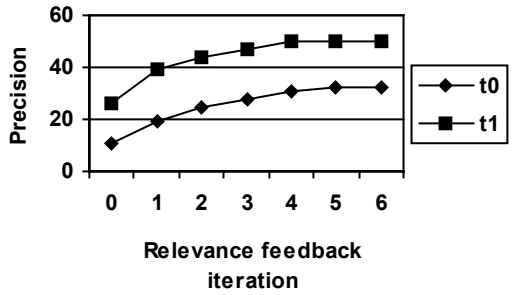

Figure 2 – Average precision of queries in two moments

As expected the retrieval accuracy is improved with the growing usage of the system, since it learns from the user in order to update the association text-image.

5. CONCLUSIONS AND FUTURE WORK

The approach presented for relevance feedback in image retrieval was based on a dependency between the short-term or session level process and the long term or concept learning process. Clearly, at the first stage of usage of the system, the influence of the first is predominant while, with time, the influence of the learned concepts is bigger.

Currently, other characteristics already incorporated in the VOIR prototype are being tested and evaluated as an extension of the discussed approach. Among these is the possibility of using more than one region in each query formulation, the use of spatial relationships between query regions, and the exploration of additional relationships between concepts in the query process.

REFERENCES

[1] Armitage, L. and Enser, P. G. B., "Analysis of user need in image archives," *Journal of Information Science*, vol. 23, no. 4, pp. 287-299, 1997

[2] Enser, P. G. B., "Query Analysis in a visual information retrieval context," *Journal of Document and Text Management*, vol. 1, no. 1, pp. 25-52, 1993

[3] Rui, Y., Huang, T. S., and Chang, S. F., "Image Retrieval: Current Techniques, Promising Directions and Open Issues," *Journal of Visual Communication and Image Representation*, vol. 10 pp. 39-62, 1999

[4] Rocchio, J., "Relevance feedback in information retrieval," in Salton, G. (ed.) *The SMART Retrieval System - Experiments in Automatic Document Processing* Prentice-Hall, 1971, pp. 313-323

[5] Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S., "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644-655, 1998

[6] Ishikawa, Y., Subramanya, R., and Faloutsos, C. MindReader : querying databases through multiple examples. Proc. 24th Int. Conf. Very Large Data Bases, VLDB. 1998. Pittsburgh, Pa. : Department of Computer Science, Carnegie-Mellon University. CMU-CS- ; 98-119.

[7] Rui, Y. and Huang, T. S., "A Novel Relevance Feedback Technique in Image Retrieval," *ACM Multimedia*, pp. 67-70, 1999

[8] Wu, L., Faloutsos, C., Sycara, K., and Payne, T. R. FALCON Feedback Adaptive Loop for Content-Based Retrieval. VLDB 2000. 2000.

[9] Lu, Y., Hu, C., Zhu, X., Zhang, H., and Yang, Q. A unified framework for semantics and feature based relevance feedback in image retrieval systems. Proc. of ACM Multimedia 2000. 31-38. 2000. Los Angeles, USA.

[10] Jing, F., Li, M., Zhang, H., and Zhang, B. Region-Based Relevance Feedback in Image Retrieval. IEEE ISCAS 2002 - Symposium on Circuits and Systems. 2002. Arizona, USA.

[11] Zhuang, Y., Yang, J., and Li, Q. A Graphic-Theoretic Model for Incremental Relevance Feedback in Image Retrieval. Proc. IEEE Int. Conf. on Image Processing 2002. 2002. New York, USA.

[12] Lee, C. S., Ma, W. Y., and Zhang, H. Information Embedding Based on User's Relevance Feedback for Image Retrieval. SPIE Photonic East. 1999. Boston, USA.

[13] Torres, J. and Parkes, A. A Modular Framework for Visual Object Information Retrieval. Izquierdo, E. Proc. of the 4th Workshop on Image Analysis for Multimedia Interactive Services. 73-76. 2003. Queen Mary, University of London.

[14] APT. Australian Pictorial Thesaurus. Council of Australian State Libraries (CASL) . 2002.

[15] Manjunath, B. S., Salembier, P., and Sikora, T., *Introduction to MPEG-7, Multimedia Content Description Interface* John Wiley & Sons Ltd., 2002,

[16] Gennari, J. H., Langley, P., and Fisher, D. H., "Models of incremental concept formation," *Artificial Intelligence*, vol. 40 pp. 11-60, 1989

[17] Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its applications to evaluating segmentation algorithms and measuring ecological statistics. Proc. IEEE 8th Int. Conf. Computer Vision. 416-423. 2001. Vancouver, Canada.