Performance boosting with three mouse clicks - Relevance feedback for CBIR

Daniel Heesch and Stefan Rüger

Department of Computing, Imperial College
180 Queen's Gate, London SW7 2BZ, England
{dh500,srueger}@doc.ic.ac.uk

Abstract. In this paper we introduce a novel relevance feedback method for content-based image retrieval and demonstrate its effectiveness using a subset of the Corel Gallery photograph collection and five low-level colour descriptors. Relevance information is translated into updated, analytically computed descriptor weights and a new query representation, and thus the system combines movement in both query and weight space. To assess the effectiveness of relevance feedback, we first determine the weight set that is optimal on average for a range of possible queries. The resulting multiple-descriptor retrieval model yields significant performance gains over all the single-descriptor models and provides the benchmark against which we measure the additional improvement through relevance feedback. We model a number of scenarios of user-system interaction that differ with respect to the precise type and the extent of relevance feedback. In all scenarios, relevance feedback leads to a significant improvement of retrieval performance suggesting that feedback-induced performance gain is a robust phenomenon. Based on a comparison of the different scenarios, we identify optimal interaction models that yield high performance gains at a low operational cost for the user. To support the proposed relevant feedback technique we developed a novel presentation paradigm that allows relevance to be treated as a continuous variable.

1 Introduction

The efficient retrieval of images based on automatically extracted image data has become of great interest with a number of application areas ranging from medical imaging to remote sensing. Although the number of commercial and research systems addressing content-based image retrieval (CBIR) is growing rapidly, the area is far from reaching maturity. Further progress is currently impeded by two fundamental problems. The first arises from the difficulty of inferring image meaning from primitive image features such as texture, colour and shape. It seems that a high-level representation of the image that speaks the same language as the user cannot be build from intrinsic image data alone but that a significant amount of world knowledge is essential. This problem which is commonly referred to as the semantic gap is quite distinct from, although

sometimes confused with, the second problem which results from the semantic ambiguity inherent in images. Even if it were possible to derive image meaning from primitive features, and thus to bridge the semantic gap, the problem of polysemy would persist for it would still not be clear *a priori* which of the many high-level representations a user has in mind when querying a database with an example image.

User-system interaction is a promising step towards tackling the above challenges for it can furnish the system with more information than is contained in the image itself. In the earlier days of CBIR, interaction was limited to the user specifying various system settings such as the particular features to be employed for retrieval. This type of interaction binds the user to a primitive level that becomes very obscure once more sophisticated image descriptors are used. While most of the commercial systems are still limited to this type of interaction, a number of ideas have been developed with the aim of more efficiently eliciting additional information from the user. The mechanisms that have emerged include *inter alia* supervised learning prior to retrieval [9], interactive region segmentation [4], and interactive image database annotation [12]. The mechanism which holds most promise of improving the efficiency of retrieval systems is relevance feedback [19], i.e. the labelling of retrieved images according to their perceived relevance. Unlike with text documents, looking at images and deciding whether they are relevant to a query constitute a relatively small mental load. A number of different techniques have been used for relevance feedback. They broadly fall into one of two categories, namely (i) query point moving and (ii) weight update [1]. The first comprises all those methods that alter the query representation [11]. The idea is to find not the best aggregation of image features but a representation of the query that renders it more similar to objects the user has marked out as relevant or, likewise, more dissimilar to non-relevant objects (e.g. [7, 14, 8]). Given a set of features, the similarity values derived for individual features are aggregated to yield an overall measure of similarity. Aggregation is often done by computing a weighted average of individual similarity values. Weight update techniques exploit relevance feedback to infer which features (and which representations thereof) best capture the user's perception of similarity. Examples include cluster analysis of the images using Kohonen's Learning Vector Quantization [22], non-parametric density estimation [11], the use of inverted files for feature selection [20], Bayesian Networks [3, 11], a statistical analysis of the feature distributions of relevant images [2], variance analysis [17] and analytic global optimisation [16].

In our system, we combine the idea of query point moving with weight update. Retrieved images are placed on the screen such that their distances to the centre represent their dissimilarities to the query image (see Figure 1). By moving an image closer to the centre or further away, the user indicates to what extent the image matches his information need. Because this match is given in terms of the chosen distance from the center and thus effectively continuous, this method of relevance feedback is likely to provide a more accurate representation of the user's idea of similarity. With the similarity function being a simple convex weight

combination of basic feature similarity functions, new weight coefficients can be computed in such a way that the user's perception of similarity with respect to the query is as close as possible to the similarities computed by the system. The similarity model chosen allows us to compute the new weights analytically using a least squared error approach.

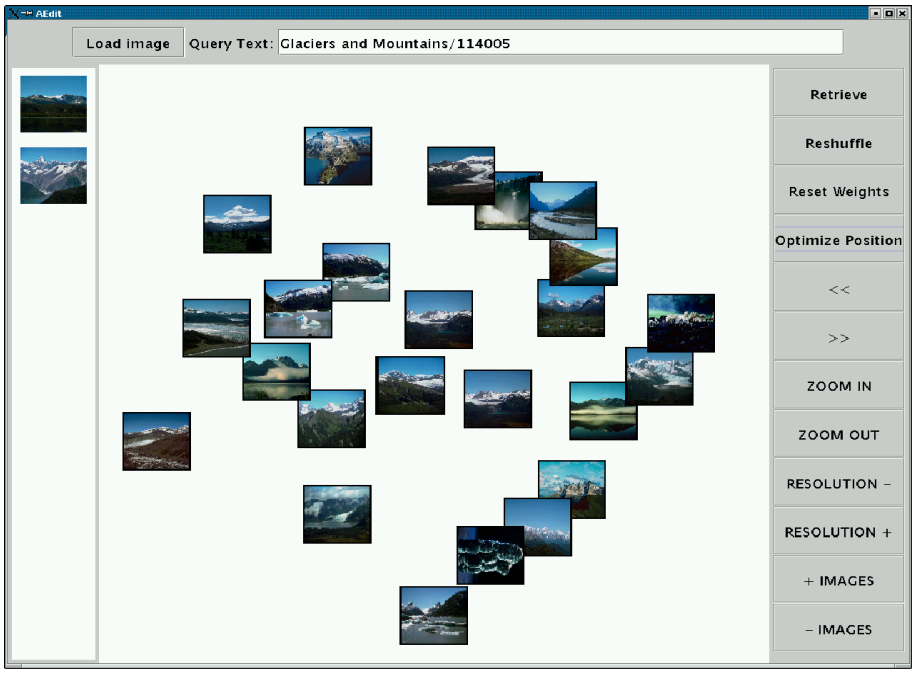

Fig. 1. Screen shot of the user interface. The query consists of two images of snow-peaked mountains, depicted at the top of the leftmost canvas. The display shows the 30 top-ranked images. The distance of a retrieved image from the centre is proportional to its dissimilarity to the query as computed by the system. Subject to this distance constraint, images are arranged such that overlap is minimized.

Operationally, the relevance feedback approach of the present study is similar to [18], where the interface consists in a projection of the image objects in feature space onto the two dimensional plane using a distance metric derived from the user's interaction with the display. Instead of querying a database with an example, the user is invited to explore the database by changing the projection operator. This is done by marking out relevant and non-relevant images as well as by moving images closer to or further away from other images. In our system, rather than exploring the database, the user queries the database with an example image with the semantics of the interaction being kept deliberately simple. While in [18] the arrangement of images on the screen is such that distances

inform about respective dissimilarities, we consider and display dissimilarity of an image with respect to a query.

A challenge pertaining to relevance feedback techniques lies in striking a balance between simplicity and efficiency. Ideally, feedback results in a significant gain in retrieval performance after a few iterations with a minimum of user operations. With this in mind, we have modelled different scenarios of relevance feedback that vary in the quantity and quality of the interaction. It turns out that the efficiency of relevance feedback is a general outcome across a broad range of interaction scenarios and, more importantly, that it is possible to define conditions under which the system rewards even the most partial feedback.

In summary, the contributions of this papers to CBIR are essentially threefold. Firstly, we describe a relevance feedback technique that is based on merging techniques of query point moving with those of weight update. Secondly, to support relevance feedback, we develop a simple yet highly expressive way of presenting the user with the retrieved images. Lastly, we model different scenarios of user-system interaction in order to formally quantify the improvement in retrieval performance that can be achieved through relevance feedback and to identify interaction models that are both user-friendly and efficient.

The paper is divided into four sections. Section 2 describes the five colour descriptors and introduces the relevance feedback technique along with the different interaction models used for evaluation. The results are presented in Section 3 and discussed in Section 4.

2 Methods

2.1 Colour descriptors

Descriptors for colour representation were chosen with the new MPEG-7 standard in mind. Both the colour structure descriptor (defined in the HMMD colour space) and the colour cluster descriptor (defined in the YCbCr colour space) are part of the MPEG-7 specification. It was also desirable to keep the set of descriptors sufficiently distinct from one another with the aim of representing different aspects of the image colour. Choosing descriptors from different colour spaces and allowing for different spatial emphasis appeared to be sensible ways to decrese representational overlap and motivated the addition of the standard HSV colour histogram and a local variant thereof.

HSV colour histogram We use a uniform quantization of the HSV colour space using 8 bins for each dimension. Two different descriptors are defined. One is a global histogram consisting of $8 \times 8 \times 8$ colour bins with loss of spatial information, while the other consists of an array of five local histograms each of size $8 \times 8 \times 8$, with the first four covering non-overlapping quarters of the image and the fifth covering the central 25% of the image. This second descriptor allows central emphasis and partial preservation of local colour information.

HMMD colour histogram The new HMMD (Hue, Min, Max, Diff) colour space, which is supported by MPEG-7, derives from the HSV and RGB spaces. The hue component is as in the HSV space, and max and min denote the maximum and minimum among the R, G, and B values, respectively. The Diff component is defined as the difference between Max and Min. Three components suffice to uniquely locate a point in the colour space and thus the space is effectively three-dimensional. Following the MPEG-7 standard, we quantize the HMMD non-uniformly into 184 bins with the three dimensions being Hue, Sum and Diff (Sum being defined as (Max+Min)/2). For details about quantization see [10]. Two descriptors are defined with respect to the HMMD colour space. The first is a standard global histogram, the second, CSD or colour structure descriptor, is described in more detail below.

Colour Structure Descriptor This descriptor lends itself well for capturing local colour structure in an image. A 8×8 structuring window is used to slide over the image. Each of the 184 bins of the HMMD histogram contains the number of window positions for which there is at least one pixel falling into the bin under consideration. This descriptor is capable of discriminating between images that have the same global colour distribution but different local colour structures. Although the number of samples in the 8×8 structuring window is kept constant (64), the spatial extent of the window differs depending on the size of the image. Thus, for larger images appropriate sub-sampling is employed to keep the total number of samples per image roughly constant. The bin values are normalized by dividing by the number of locations of the structuring window and fall in the range $[0.0, 1.0]$. (see [10] for details)

Colour Cluster Descriptor The colour cluster descriptor is similar to the dominant colour descriptor defined as part of the MPEG-7 standard [10]. It exploits the fact that for most images across a wide range of applications, a small set of colours are sufficient to represent the global colour content. The colour space chosen is the perceptually uniform YCbCr colour space where

$$\begin{aligned} Y &= 0.299\ R + 0.587\, G + 0.114\ B \\ Cb &= -0.169\, R - 0.331\, G + 0.500\ B \\ Cr &= 0.500\ R - 0.419\, G - 0.081\ B \end{aligned}$$

With the issue of extraction efficiency in mind, the representative colours are determined by first computing a traditional colour histogram and then successively merging those non-empty bins which are closest in the colour space. An alternative with greater complexity consists in performing agglomerative clustering starting at the level of individual pixels [5].

2.2 Combination of image descriptors

Descriptors are integrated such that the overall similarity between two images Q and T is given by a convex combination

$$S(Q,T) = \sum_d w_d \sigma_d(Q,T) \tag{1}$$

of the similarity values calculated for each descriptor. Here $\sigma_d(Q,T)$ denotes the similarity for feature d of images Q and T using a possibly feature-specific similarity function and weighted by a factor $w_d \in [0,1]$ with $0 \leq w_d \leq 1$ and $\sum_d w_d = 1$. For all but the colour cluster descriptor we use as a similarity metric the l_1 norm which has successfully been employed in the past for histogram comparisons.

$$||h_1 - h_2||_1 = \sum_{i=0}^{N-1} |h_1(i) - h_2(i)|$$

where h_1 and h_2 are the two histograms being compared. For comparing two sets of colour clusters we compute the earth mover's distance as used in [15].

We allow the user to add more objects to the original query. With $Q = \{Q_1, Q_2, \ldots Q_n\}$ defining the set of images in a query, we define the similarity between a query and a database object as

$$S(Q,T) = \sum_d w_d \frac{1}{n} \sum_i \sigma_d(Q_i,T) \tag{2}$$

2.3 Evaluation

It needs little emphasis that claims regarding improved retrieval performance can only be validated using formal evaluation methods. Although evaluation is arguably more problematic in image retrieval as a result of semantic ambiguity, evaluation techniques from information retrieval may still be used profitably when applied with care. For the present study we deploy the two measures *recall* and *precision* to measure retrieval performance for *category searches* [11, 3]. The image collection used for testing and training is derived from the Corel Gallery 380,000 package, which contains some 30,000 photographs, sorted into 450 categories. From the set of categories containing more than 40 images, we randomly selected 47 categories and for each category 40 images leaving us with a total of 1880 images. This image corpus is then split into two non-overlapping training and test sets. Each of these sets contains 20 images of each of the 47 categories[1]. To find an initial weight set for the relevance feedback runs, we use each object in the training set as a query and measure retrieval performance when retrieving from the remaining set of training images. We then average

[1] A list of these categories and associated images for both the testing and training set can be found at http://rowan.doc.ic.ac.uk/corel.

performance over all queries and repeat the procedure for different weight sets. The weight set for which the average performance is maximal is then chosen as the initial weight set. The relevance feedback runs are carried out with all objects of the test set used as a query. Retrieval is from the remaining set of test images. For each query, we determine the baseline performance using the initial weight set determined on the training set and then measure the gain (or loss) in retrieval performance after repeatedly providing relevance feedback and thereby altering the weights. Performance is measured by first determining the precision-against-recall values as described in [21] and then deriving from these the *mean average precision* as a more concise measure.

2.4 Relevance feedback

The mean average precision value for the multi-descriptor model is derived by averaging over all queries of the training set and it is very likely that the optimum weight set for *any particular* query is different from the weight set derived in 3.1. Furthermore, the discrepancy between *query* optimum and *query set* optimum is likely to be greater for heterogenous image collections (e.g. Corel Gallery). In the context of sketch retrieval, previous experiments in which we established the optimum weight sets for a number of smaller subsets of a sketch database confirm that the optimum does vary substantially between queries [6]. It is hoped that relevance feedback allows us to move from the *query set* optimum to each individual *query* optimum within a few iterations.

With five descriptor weights and the convexity assumption, our system has effectively four degrees of freedom which can be exploited through relevance feedback. Our retrieval system plots thumbnails of retrieved images $T_1, T_2, \ldots$ such that their respective distance from the centre of the screen is proportional to the dissimilarity $1 - S(Q, T_i)$ of thumbnail T_i to the query Q. Using this semantics of thumbnail location, the user can provide relevance feedback by moving thumbnails closer to the centre (indicating greater relevance than the system predicted) or further away (indicating less relevance). As a shortcut, clicking once on a thumbnail marks the image as highly relevant and places it in the centre, while double-clicking marks it as highly non-relevant and places it in the periphery. Effectively, the user provides a real-valued vector of distances $D_u(Q, T_i)$ which, in general, differ from the distances

$$D_s(Q, T_i) = 1 - \sum_d w_d \sigma_d(Q, T_i) \tag{3}$$

which the system computes using the set of weights w_d. The sum of squared errors

$$\begin{aligned} \mathrm{SSE}(w) &= \sum_{i=1}^{N} [D_s(Q, T_i) - D_u(Q, T_i)]^2 \\ &= \sum_{i=1}^{N} \Big[1 - \sum_d w_d \sigma_d(Q, T_i) - D_u(Q, T_i)\Big]^2 \end{aligned} \tag{4}$$

gives rise to an optimisation problem for the weights w_d such that (4) is minimized under the constraint of convexity. Using one Lagrangian multiplier we arrive at an analytical solution w' for the weight set which changes the similarity function. We get a different ranking of images in the database and, with (3), a new layout for the new set of top-retrieved images on the screen.

2.5 Modelling user-system interaction

We can envisage an "ideal" user who provides maximum feedback by interacting with each of the displayed images. This is a plausible scenario when the number of displayed images is small but is of little practical interest once the number of images increase, in which case we may assume that the user will want to limit her or his feedback to a subset of images. To evaluate the effect of relevance feedback on retrieval performance and, more specifically, to identify efficient ways of providing feedback, we model four different interaction scenarios. All four scenarios are based on the assumption that the similarity judgment of the user is fixed. In particular, an image is relevant if it is in the same category as the query and it is not relevant otherwise. Each scenario now defines a different set $\mathcal{S}$ of images with which the user interacts as follows:

- W$\pm$: Indifferent selection, the user interacts with a random subset of displayed images
- W$-$: Negative selection, the user only interacts with a random subset of irrelevant displayed images
- W$+$: Positive selection, the user only interacts with a random subset of relevant displayed images
- QW$+$: Same as Scenario W$+$, except that in addition to changing the distances between relevant objects and the query, the relevant objects are included in the query

For all scenarios we display the top 30 images as seen in Figure 1. The four models differ in the quality of user feedback. To model the extent to which feedback is given, we vary the maximum number (s) of images the user may interact with. Note that in Scenarios W$-$, W$+$ and QW$+$, it is $|\mathcal{S}| \leq s$ since there might not be enough relevant or irrelevant images on display, while in Scenario W$\pm$ it is $|\mathcal{S}| = s$.

3 Results

3.1 Tuning the system for optimal *on average* performance

We evaluated retrieval performance for each of the five descriptors individually and used a genetic algorithm to find the best combination of descriptors (i.e. the weights set that maximizes *on average* performance on the training set as detailed in the section 2.3. The precision-against-recall graphs for all five single-descriptor models and the best multi-descriptor model are depicted in Figure

2. For subsequent relevance feedback experiments, we therefore chose the initial weights as follows: HSVN (0.526), HSVF (0.364), CSD (0.032), HMMD (0.076), Cluster (0.002).

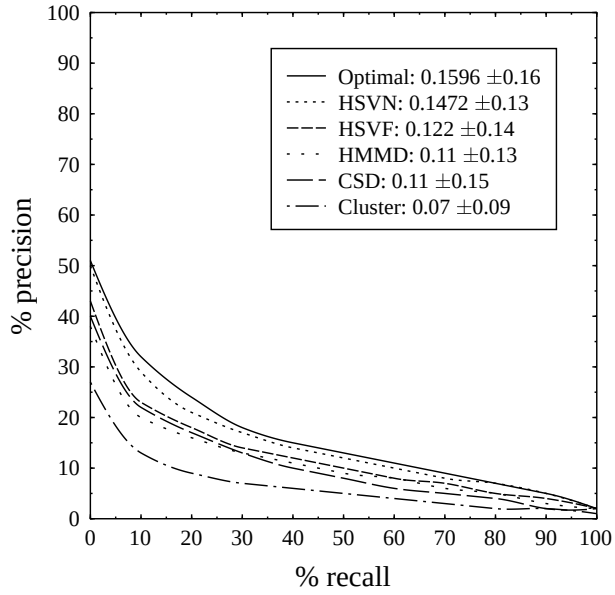

Fig. 2. Precision-against-recall graphs for the five single descriptor models and the multi-descriptor model using the optimum combination of weights: 0.526 (HSVN = global HSV histogram), 0.364 (HSVF = local HSV histogram), 0.032 (CSD = Colour Structure Descriptor), 0.076 (HMMD Descriptor) and 0.002 (Colour Cluster Descriptor)). Performance of the best multi-descriptor model improved significantly over that of the best single-descriptor model (paired t-test, $p < 0.05$) and is used as the baseline for relevance feedback evaluation.

3.2 Evaluation of relevance feedback

Our principal interest lies in investigating the possibility of using relevance feedback to increase retrieval performance beyond the level of the multiple-descriptor model. We here show the results for the four scenarios introduced in 2.5. The results are given in terms of *absolute* percentage differences: an increase of, say, 10% of the mean average precision through relevance feedback refers to a boost from 0.16 (baseline model with best constant weight setting) to 0.26.

Figure 3 summarizes the results for scenario W±, in which relevance feedback is given on both relevant and non-relevant images with the maximum number of images on which relevance feedback is given varying between 3 and 10. As is immediately evident from the graphs, gains are higher when relevance feedback

is more complete. The difference becomes less marked, however, after a few iterations with all gain curves levelling off after around four iterations.

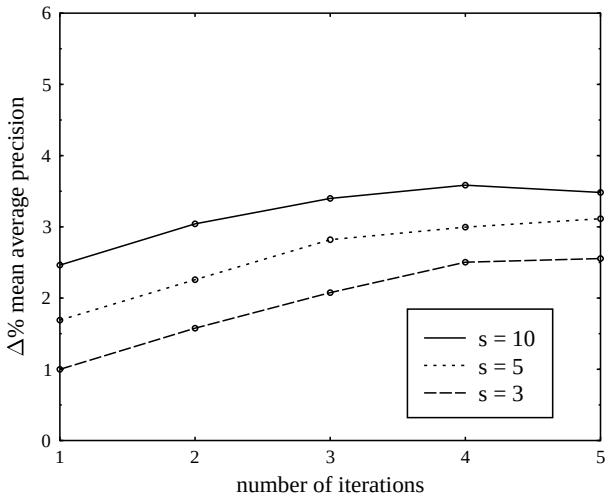

Fig. 3. Efficiency of relevance feedback under Scenario W± with varying s

Scenarios W− and W+ are designed to investigate the impact of negative and positive feedback, respectively, on retrieval performance. Intuitively, the gain should be considerably lower under Scenario W− than either scenario Scenario W± or Scenario W+ for it appears very difficult to infer what an object *is* from information about what it is not. This is precisely what Figure 4 tells us. Relevance feedback does improve performance but this improvement, at least over the first few iterations, is marginal. As before, the efficiency of relevance feedback increases with the extent of the interaction (s) and at least for the first iteration this relationship is nearly linear. Note that Scenario W± only differs from Scenario W− by allowing positive feedback and, thus, the difference between the results from Scenario W± and Scenario W− may in some sense be attributed to the additional effect of positive feedback. Comparison of the two scenarios already suggests, therefore, that positive feedback alone may be superior to either alternatives.

Figure 5 now shows the result for positive feedback. Already after one iteration, gains lie as high as 3% to 4% depending on the extent of the interaction. Unlike in the previous two scenarios, continued feedback does not provide further improvements which, lacking additional data, may at least suggest that the system has found the optimal weight set after one or two iteration. Note also that the results do not vary greatly with s. This is likely to result from the fact that quite often the number of relevant images displayed is low (typically below 10) and thus the *effective* extent of the interaction may often be the same no matter whether s is 5 or 10.

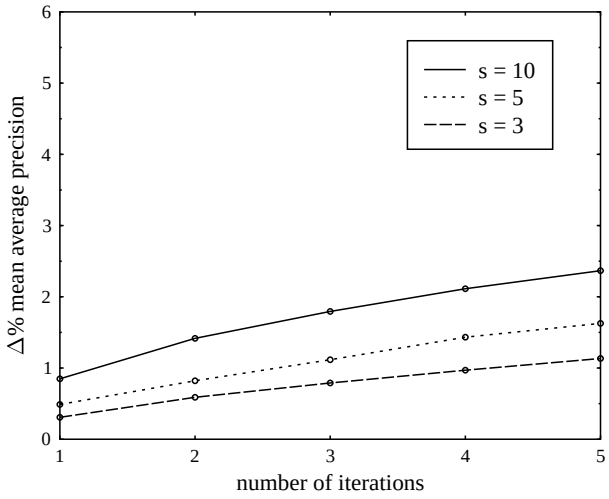

Fig. 4. Efficiency of relevance feedback under Scenario W− with varying s

One needs to keep in mind that our models assume a starting weight set which is already optimized. If a random initial weight set is chosen, feedback-induced performance gains are higher (results not shown). Given our shortcut for positive feedback (one mouseclick per image), three mouseclicks under Scenario W+ achieve a considerable performance boost beyond the level of the globally optimized multiple-descriptor model.

Let us finally turn to Scenario QW+ in which the original query is expanded by those images marked as relevant (Figure 6). After only one iteration, performance has risen by 10.5-11%. As has already been observed for Scenario W+ and for the same reason, the performance gain does not greatly depend on s. Unlike in Scenario W+ however, continuous feedback results in additional improvement which may suggest that the addition of new images to the query allows the system to widen its "catchment area". Scenario QW+ thus achieves a substantial *and* instantaneous performance boost which is unmatched by any other scenario considered.

The results after *one* iteration of relevance feedback are summarized in Table 1.

4 Discussion

We have described and evaluated a novel relevance feedback method for CBIR that combines analytical weight update with query expansion. Evaluation was done automatically by modelling a number of user-system interaction scenarios. Our results show that feedback-induced performance gains over an already optimized multiple-descriptor model occur under a wide range of interaction scenarios. They also make very clear that some scenarios are decidedly better than

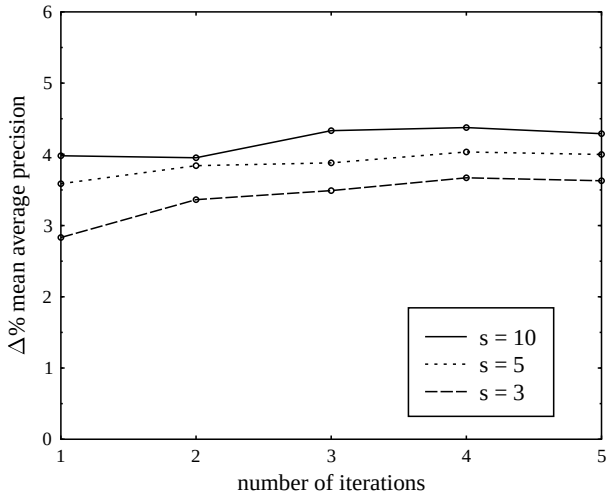

Fig. 5. Efficiency of relevance feedback under Scenario W+ with varying s

Scenario \ clicks	10	5	3
W−	0.9	0.5	0.3
W±	2.5	1.8	1.0
W+	4.0	3.6	2.9
QW+	12.8	12.0	10.5

Table 1. The gain (in $\Delta\%$) in mean average precision for the four different scenarios, a constant display size of 30 and varying s (= number of mouse clicks). By far the highest values are achieved under Scenario QW+.

others. Specifically, the data support the view that, for any level of interaction (i.e. the number of images on which feedback is given), positive feedback results in greater and more immediate improvement than either negative or mixed feedback. We show that through query expansion, the effect of positive feedback gets multiplied considerably and that this combination allows substantial gains in retrieval performance after only one iteration with a minimum of user-interaction (three mouse clicks).

As argued in the introduction, semantic ambiguity constitutes a challenge for CBIR which may be overcome through techniques such as relevance feedback. In this paper, by deciding *a priori* whether or not an image is relevant to a given query, it appears, however,as if we effectively ignore the issue of image polysemy. It is certainly true that we do enforce a particular "reading" of an image by assigning it to one particular category. So the image of a yellow flower may happen to be in the category "Yellow" but not in the category "Flower" and would therefore be considered irrelevant with respect to any query image from the latter category. We can conceive a user who would deem it relevant,

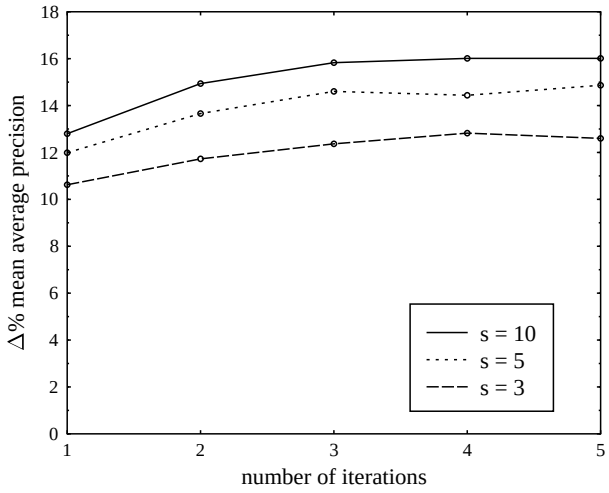

Fig. 6. Efficiency of relevance feedback under Scenario QW+ with varying s

however, and this user would effectively not be modelled by our system. The important point is that the user we *do* model is essentially arbitrary. The categorization of images (and thus the fixation of image meaning) has not been done such that certain strengths of the system get emphasized (as is sometimes done, whether deliberately or not [13]). Our categories are chosen at random from the total number of categories and comprise highly abstract categories which one may not expect a system operating on low-level features to retrieve successfully. This suggests that the efficiency of relevance feedback demonstrated in this paper is relatively independent of the user. Meta-experiments with different categorizations would provide a way of assessing more precisely the degree of user-independence. Additional evidence that the results presented here are of a generic nature comes from retrieval experiments (results unpublished) we performed on a sketch collection of similar size as the one used for this paper using a small set of low-level shape descriptors.

In its early years CBIR has been largely concerned with finding better ways to represent particular image features. Our results strongly support the view that further progress can be made by addressing the questions of how to integrate evidence from multiple of these generally low-level feature representations, and more importantly, how to elicit and utilize user-specific relevance information. Our results are all the more encouraging as we have thus far considered only one image feature, namely colour. We may reasonably expect that with texture and shape representations being added to the same retrieval model, the positive effects of relevance feedback will be even greater.

References

1. Y-S Choi, D Kim, and R Krishnapuram. Relevance feedback for content-based retrieval using the Choquet integral. In *Proc of the IEEE International Conference on Multimedia and Expo (II)*, 2000.
2. G Ciocca and R Schettini. A relevance feedback mechanism for content-based image retrieval. *Information Processing and Management*, 35(5):605–632, 1999.
3. I J Cox, M L Miller, T P Minka, T V Papathomas, and P N Yianilos. The Bayesian image retrieval system, pichunter. *IEEE Transactions on Image Processing*, 9(1):20–38, 2000.
4. D Daneels, D Campenhout, W Niblack, W Equitz, R Barber, E Bellon, and F Fierens. Interactive outlining: an improved approach using active contours. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1993.
5. Y Deng and B S Manjunath. An efficient color representation for image retrieval. *IEEE Transactions on Image Processing*, 10:140–147, 2001.
6. D Heesch and S Rüger. Combining features for content-based sketch retrieval — a comparative evaluation of retrieval performance. In *Proceedings of the European Colloquium on IR Research 2002*, Berlin, 2002. LNCS, Springer.
7. Y Ishikawa, R Subramanya, and C Faloutsos. MindReader: Querying databases through multiple examples. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 218–227, 24–27 1998.
8. A Lelescu, O Wolfson, and B Xu. Approximate retrieval from multimedia databases using relevance feedback. In *SPIRE/CRIWG*, pages 215–223, 1999.
9. W Y Ma and B S Manjunath. Texture features and learning similarity. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 425–430, 1996.
10. B S Manjunath and J-R Ohm. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11:703–715, 2001.
11. C Meilhac and C Nastar. Relevance feedback and category search in image databases. In *Proc. IEEE Int. Conf. Multimedia Comp. and Syst.*, pages 512–517, 1999.
12. T P Minka and R W Picard. Interactive learning using a society of models. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 447–452, 1996.
13. H Mueller, S Marchand-Maillet, and T Pun. The truth about Corel - evaluation in image retrieval. In *Proceedings of CIVR*, pages 38–49, 2002.
14. K Porkaew, M Ortega, and S Mehrotra. Query reformulation for content based multimedia retrieval in MARS. In *ICMCS, Vol. 2*, pages 747–751, 1999.
15. Y Rubner and L J Guibas. The earth mover's distance, multi-dimensional scaling and color-based image retrieval. In *Proceedings of the APRA Image Understanding Workshop*, pages 661–668, 1997.
16. Y Rui and T S Huang. A novel relevance feedback technique in image retrieval. In *ACM Multimedia (2)*, pages 67–70, 1999.
17. Y Rui, T S Huang, and S Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.
18. S Santini, A Gupta, and R Jain. Emergent semantics through interaction in image databases. *IEEE transactions on knowledge and data engineering*, 13(3):337–351, 2001.
19. 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 Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1379, 2000.

20. D McG Squire, W Müller, H Müller, and T Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(13–14):1193–1198, 2000.
21. E M Voorhees and D Harman. Overview of the eigth Text REtrieval Conference (TREC-8). In *Proc. TREC*, pages 1–33 and A.17 – A.18, 1999.
22. M E J Wood, N W Campbell, and B T Thomas. Iterative refinement by relevance feedback in content-based digital image retrieval. In *ACM Multimedia 98*, pages 13–20, Bristol, UK, 1998. ACM.

Acknowledgements: This work was partially supported by the EPSRC, UK.