A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems

Ye Lu[1*], Chunhui Hu[2], Xingquan Zhu[3*], HongJiang Zhang[2], Qiang Yang[1*]

1 School of Computing Science
Simon Fraser University
Burnaby, B.C., Canada, V5A1S6
{yel,qyang}@cs.sfu.ca

2 Microsoft Research China
5F, Beijing Sigma Center
Beijing 100080, China
{i-chhu,hjzhang}@microsoft.com

3 Department of Computer Science
Fudan University
Shanghai 200433, China
980015@fudan.edu.cn

ABSTRACT

The relevance feedback approach to image retrieval is a powerful technique and has been an active research direction for the past few years. Various ad hoc parameter estimation techniques have been proposed for relevance feedback. In addition, methods that perform optimization on multi-level image content model have been formulated. However, these methods only perform relevance feedback on the low-level image features and fail to address the images' semantic content. In this paper, we propose a relevance feedback technique, *iFind*, to take advantage of the semantic contents of the images in addition to the low-level features. By forming a semantic network on top of the keyword association on the images, we are able to accurately deduce and utilize the images' semantic contents for retrieval purposes. The accuracy and effectiveness of our method is demonstrated with experimental results on real-world image collections.

Keywords

relevance feedback, image semantics, image retrieval, multimedia database.

1. INTRODUCTION

With the increasing availability of digital images, automatic image retrieval tools provide an efficient means for users to navigate through them. Even though traditional methods allow the user to post queries and obtain results, the retrieval accuracy is severely limited because of the inherent complexity of the images for users' to describe exactly. The more recent relevance feedback approach, on the other hand, reduces the needs for a user to provide accurate initial queries by estimating the user's ideal query using the positive and negative examples given by the user.

The current relevance feedback based systems estimate the ideal query parameters on only the low-level image features such as color, texture, and shape. These systems work well if the feature vectors can capture the essence of the query. For example, if the user is searching for an image with complex textures having a particular combination of colors, this query would be extremely difficult to describe but can be reasonably represented by a combination of color and texture features. Therefore, with a few positive and negative examples, the relevance feedback system will be able to return reasonably accurate results. On the other hand, if the user is searching for a specific object that cannot be sufficiently represented by combinations of available feature vectors, these relevance feedback systems will not return many relevant results even with a large number of user feedbacks.

To address the limitations of the current relevance feedback systems, we propose a framework that performs relevance feedback on both the images' semantic contents represented by keywords and the low-level feature vectors such as color, texture, and shape. The contribution of our work is twofold. First, it introduces a method to construct a semantic network on top of an image database and uses a simple machine learning technique to learn from user queries and feedbacks to further improve this semantic network. In addition, we propose a framework in which semantic and low-level feature based relevance feedback can be seamlessly integrated.

This paper is organized as follows. In Section 2, we will provide an overview of the current state of the art relevance feedback systems. In Section 3, we will present the details of our work. Section 4 will describe the *iFind* image retrieval system that we have implemented based on the proposed method and provide experimental evaluations showing its effectiveness in image retrieval. Concluding remarks will be given in Section 5.

2. RELATED WORK

---

* This work was performed at Microsoft Research China.

One of the most popular models used in information retrieval is the vector model [1, 8, 9]. Various effective retrieval techniques have been developed for this model and among them is the method of relevance feedback. Most of the previous relevance feedback research can be classified into two approaches: query point movement and re-weighting [3].

The query point movement method essentially tries to improve the estimate of the "ideal query point" by moving it towards good examples point and away from bad example points. The frequently used technique to iteratively improve this estimation is the Rocchio's formula given below for sets of relevant documents D'_R and non-relevant documents D'_N given by the user.

$$Q' = \alpha Q + \beta\left(\frac{1}{N_{R'}}\sum_{i \in D'_R} D_i\right) - \gamma\left(\frac{1}{N_{N'}}\sum_{i \in D'_N} D_i\right) \quad (1)$$

where α, β, and γ are suitable constants; $N_{R'}$ and $N_{N'}$ are the number of documents in D'_R and D'_N respectively. This technique is implemented in the MARS system [6]. Experiments show that the retrieval performance can be improved considerably by using relevance feedback [1, 8, 9].

The central idea behind the re-weighting method is very simple and intuitive. The MARS system mentioned above implements a slight refinement to the re-weighting method call the standard deviation method [6]. Since each image is represented by an N dimensional feature vector, we can view it as a point in an N dimensional space. Therefore, if the variance of the good examples is high along a principle axis j, then we can deduce that the values on this axis is not very relevant to the input query so that we assign a low weight w_j on it. Therefore, the inverse of the standard deviation of the j^{th} feature values in the feature matrix is used as the basic idea to update the weight w_j.

Recently, more computationally robust methods that perform global optimization have been proposed. The MindReader retrieval system designed by Ishikawa et al. [3] formulates a minimization problem on the parameter estimating process. Unlike traditional retrieval systems whose distance function can be represented by ellipses aligned with the coordinate axis, the MindReader system proposed a distance function that is not necessarily aligned with the coordinate axis. Therefore, it allows for correlations between attributes in addition to different weights on each component. A further improvement over this approach is given by Rui and Huang [7]. In their CBIR system, it not only formulates the optimization problem but also takes into account the multi-level image model.

All the approaches described above perform relevance feedback at the low-level feature vector level, but failed to take into account the actual semantics for the images themselves. The inherent problem with these approaches is that the low-level features are often not as powerful in representing complete semantic content of images as keywords in representing text documents. In other words, applying the relevance feedback approaches used in text information retrieval technologies to low-level feature based image retrieval will not be as successful as in text document retrieval. In viewing this, there have been efforts on incorporating semantics in relevance feedback for image retrieval. The framework proposed in [4] attempted to embed semantic information into a low-level feature based image retrieval process using a correlation matrix. In this effective framework, semantic relevance between image clusters is learnt from user's feedback and used to improve the retrieval performance. As we shall show later, our proposed method integrates both semantics and low-level features into the relevance feedback process in a new way. Only when the semantic information is not available, our method is reduced to one of the previously described low-level feedback approaches as a special case.

3. THE PROPOSED METHOD

There are two different modes of user interactions involved in typical retrieval systems. In one case, the user types in a list of keywords representing the semantic contents of the desired images. In the other case, the user provides a set of examples images as the input and the retrieval system will try to retrieve other similar images. In most image retrieval systems, these two modes of interaction are mutually exclusive. We argue that combining these two approaches and allow them to benefit from each other yields a great deal of advantage in terms of both retrieval accuracy and ease of use of the system.

In this section, we describe a method to construct a semantic network from an image database and present a simple machine learning algorithm to iteratively improve the system's performance over time. In addition, we describe a framework in which the previously constructed semantic network can be seamlessly integrated with low-level feature vector based relevance feedback.

3.1 Semantic Network

The semantic network is represented by a set of keywords having links to the images in the database. Weights are assigned to each individual link. This representation is shown pictorially as follows.

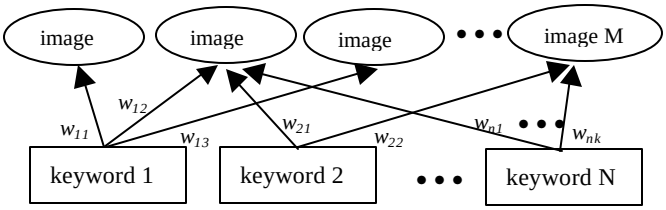

Figure 1: Semantic network

The links between the keywords and images provide structure for the network. The degree of relevance of the keywords to the associated images' semantic content is represented as the weight

on each link. It is clear that an image can be associated with multiple keywords, each of which with a different degree of relevance. Keyword associations may not be available at the beginning. There are several ways to obtain keyword associations. The first method is to simply manually label images. This method may be expensive and time consuming. To reduce the cost of manual labeling, we utilize the Internet and its countless number of users. One possible way to do that may be to implement a crawler to go to different websites to download images. We store the information such as the file name and the ALT tag string within the IMAGE tags of the HTML files as keywords associated with the downloaded image. Also, the link string and the title of the page may be somewhat related to the image. We assign weights to these keyword links according to their relevance. Heuristically, we list this information in the order of descending relevance: the link string, the ALT tag string, the file name, and the title of the page. Another approach to incorporate additional keywords into the system would be to utilize the user's input queries. Whenever the user feeds back a set of image being relevant to the current query, we add the input keywords into the system and link them with these images. In addition, since the user tells us that these images are relevant, we can confidently assign a large weight on each of the newly created links. This effectively suggests a very simple voting scheme for updating the semantic network in which the keywords with a majority of user consensus will emerge as the dominant representation of the semantic content of their associated images.

3.2 Semantic Based Relevance Feedback

Semantic based relevance feedback can be performed relatively easily compared to its low-level feature counterpart. The basic idea behind it is a simple voting scheme to update the weights w_{ij} associated with each link shown in Figure 1 without any user intervention. The weight updating process is described below.

1. Initialize all weight w_{ij} to 1. That is, every keyword has the same importance.
2. Collect the user query and the positive and negative feedback examples.
3. For each keyword in the input query, check to see if any of them is not in the keyword database. If so, add them into the database without creating any links.
4. For each positive example, check to see if any query keyword is not linked to it. If so, create a link with weight 1 from each missing keyword to this image. For all other keywords that are already linked to this image, increment the weight by 1.
5. For each negative example, check to see if any query keyword is linked with it. If so, set the new weight $w_{ij}'=w_{ij}/4$. If the weight w_{ij} on any link is less than 1, delete that link.

It can be easily seen that as more queries are inputted into the system, the system is able to expand its vocabulary. Also, through this voting process, the keywords that represent the actual semantic content of each image will receive a large weight.

The weight w_{ij} associated on each link of a keyword represents the degree of relevance in which this keyword describes the linked image's semantic content. For retrieval purposes, we need to consider another aspect. The importance of keywords that have links spreading over a large number of images in the database should be penalized. Therefore, we suggest the relevance factor r_k of the k^{th} keyword association be computed as follows.

$$r_k = w_k \left(\log_2 \frac{M}{d_i} + 1\right) \tag{2}$$

where M is the total number of images in the database, $w_k = w_{mn}$ if $m = i$ and 0 otherwise, and d_i is the number of links i^{th} keyword has.

3.3 Integration with Low-Level Feature Based Relevance Feedback

Since [7] summarized a general framework in which all the other low-level feature based relevance feedback methods discussed in Section 2 can be viewed as its special cases, in this section, we show how the semantic relevance feedback method can be seamlessly integrated with it.

To expand the framework summarized in [7] to include semantic feedback, notice that the inputs to it are a query vector q_i associated with the i^{th} feature, an N element vector $p=[p_1,...p_N]$ that represents the degree of relevance for each of the N input training samples, and a set of N training vectors x_{ni} for each feature i. As shown in [7], the ideal query vector q_i^* for feature i is the weighted average of the training samples for feature i given by

$$q_i^{T*} = \frac{\boldsymbol{p}^T X_i}{\sum_{n=1}^{N} \boldsymbol{p}_n} \tag{3}$$

where X_i is the $N \times K_i$ training sample matrix for feature i, obtained by stacking the N training vectors x_{ni} into a matrix. The optimal weight matrix W_i^* is given by

$$W_i^* = (\det(C_i))^{\frac{1}{K_i}} C_i^{-1} \tag{4}$$

where C_i is the weighted covariance matrix of X_i. That is

$$C_{i_{rs}} = \frac{\sum_{n=1}^{N} \boldsymbol{p}_n (x_{nir} - q_{ir})(x_{nis} - q_{is})}{\sum_{n=1}^{N} \boldsymbol{p}_n} \qquad r,s = 1,\ldots K_i \tag{5}$$

We can see from the above equations that the critical inputs into the system are x_{ni} and p. Initially, the user inputs these data to the system. However, we can eliminate this first step by automatically providing the system with this initial data. This is done by searching the semantic network for keywords that appear in the input query. From these keywords, we can follow the links to obtain the set of training images (duplicate images are removed). The vectors x_{ni} can be computed easily from the training set. To

compute the degree of relevance vector $\boldsymbol{p}$, we can use the following formula.

$$\boldsymbol{p}_i = \boldsymbol{a}^M \sum_{j=1}^{M} r_{ij} \tag{6}$$

where M is the number of query keywords linked to the training image i, r_{jk} is the relevance factor of the j^{th} keyword associated with image i, and $\boldsymbol{a} > 1$ is a suitable constant. We can see that the degree of relevance of the i^{th} image increases exponentially with the number of query keywords linked to it. In the current implementation of our system, we have experimentally determined that setting $\boldsymbol{a}$ to 2.5 gives the best result.

To incorporate the low-level feature based feedback and ranking results into high-level semantic feedback and ranking, we define a unified distance metric function G_j to measure the relevance of any image j within the image database in terms of both semantic and low-level feature content. The function G_j is defined using a modified form of the Rocchio's formula as follows.

$$G_j = \log(1+p_j)D_j + b\left\{\frac{1}{N_R}\sum_{k\in N_R}\left[\left(1+\frac{I_1}{A_1}\right)S_{jk}\right]\right\} - g\left\{\frac{1}{N_N}\sum_{k\in N_N}\left[\left(1+\frac{I_2}{A_2}\right)S_{jk}\right]\right\} \tag{7}$$

where D_j is the distance score computed by the low-level feedback in [7], N_R and N_N are the number of positive and negative feedbacks respectively, I_1 is the number of distinct keywords in common between the image j and all the positive feedback images, I_2 is the number of distinct keywords in common between the image j and all the negative feedback images, A_1 and A_2 are the total number of distinct keywords associated with all the positive and negative feedback images respectively, and finally S_{ij} is simply the Euclidean distance of the low-level features between the images i and j. We have replaced the first parameter $\boldsymbol{a}$ in Rocchio's formula with the logarithm of the degree of relevance of the j^{th} image. The other two parameters $\boldsymbol{b}$ and $\boldsymbol{g}$ are assigned a value of 1.0 in our current implementation of the system for the sack of simplicity. However, other values can be given to emphasize the weighting difference between the last two terms.

Using the method described above, we can perform the combined relevance feedback as follows.

1. Collect the user query keywords
2. Use the above method to compute x_{ni} and $\boldsymbol{p}$ and input them into the low-level feature relevance feedback component to obtain the initial query results.
3. Collect positive and negative feedbacks from the user
4. Update the semantic network with the method given in section 3.2
5. Update the weights of the low-level feature based component using the methods discussed in [7]
6. Compute the new x_{ni} and $\boldsymbol{p}$ and input into the low-level feedback component
7. Compute the ranking score for each image using equation 7 and sort the results.
8. Show new results and go to step 3

Usually the values of x_{ni} are computed beforehand in a pre-processing step. We can see that using this approach, our system learns from the user's feedback both semantically and in a feature based manner. In addition, it can be easily seen that our method degenerates into the method of Rui and Huang [7] when no semantic information is available. We will show in the next section how our system deals with input queries that have no associated images from the semantic network. Also, next section will present some experimental results to confirm the effectiveness of this approach.

3.4 New Image Registration

Adding new images into the database is a very common operation under many circumstances. For retrieval systems that entirely rely on low-level image features, adding new images simply involves extracting various feature vectors for the set of new images. However, since our system utilizes keywords to represent the images' semantic contents, the semantic contents of the new images have to be labeled either manually or automatically. In this section, we present a technique to perform automatic labeling of new images.

In paper [5], a method was presented which automatically classify images into only two categories, indoor and outdoor, based on both text information and low-level feature. There is currently no algorithm available to automatically determine the semantic content of arbitrary images accurately. We implemented a scheme to automatically label the new images by guessing their semantic contents using low-level features. The following is a simple algorithm to achieve this goal.

1. For each category in the database, compute the representative feature vectors by determining the centroid of all images within this category.
2. For each category in the database, find the set of representative keywords by examining the keyword association of each image in this category. The top N keywords with largest weight whose combined weight does not exceed a previously determined threshold t are selected and added into the list the representative keywords. The value of the threshold t is set of 40% of the total weight as discussed in section 4.
3. For each new image, compare its low-level feature vectors against the representative feature vectors of each category. The images are labeled with the set of representative keywords from the closest matching category with an initial weight of 1.0 on each keyword.

Because the low-level features are not enough to present the images' semantics, some or even all of the automatically labeled keywords will inevitably be inaccurate. However, through user queries and feedbacks, semantically accurate keywords labels will emerge.

Anther problem related to automatic labeling of new images is the automatic classification of these images into predefined categories. We solve this problem with the following algorithm.

1. Put the automatically labeled new images into a special "unknown" category.
2. At regular intervals, check every image in this category to see if any keyword association has received a weight greater than a threshold x. If so, extract the top N keywords whose combined weight does not exceed the threshold t.
3. For each image with extracted keywords, compare the extracted keywords with the list of representative keywords from each category. Assigned each image to the closest matching category. If none of the available categories result in a meaningful match, leave this image in the "unknown" category.

The keyword list comparison function used in step 3 of the above algorithm can take several forms. The ideal function would take into account the semantic relationship of keywords in one list with those of the other list. However, for the sake of simplicity, our system only checks for the existence of keywords from the extracted keyword list in the list of representative keywords.

4. EXPERIMENTAL RESULTS

We have presented a framework in which semantic and low-level feature based feedback can work together to achieve greater retrieval accuracy. In this section, we will describe the image retrieval system *iFind* that we have implemented using this framework and show some experimental results.

4.1 The *iFind* Retrieval System

The *iFind* image retrieval system implements the framework discussed in this paper. It is a web based retrieval system in which multiple users can perform retrieval tasks simultaneously at any given time.

The *iFind* system supports three modes of interaction: keyword based search, search by example images, as well as browsing the entire image database using a pre-defined category hierarchy. The main user interface is shown in Figure 2.

When the user enters a keyword-based query, the system invokes the combined relevance feedback mechanism discussed in Section 3.3. The result page is shown in Figure 3.

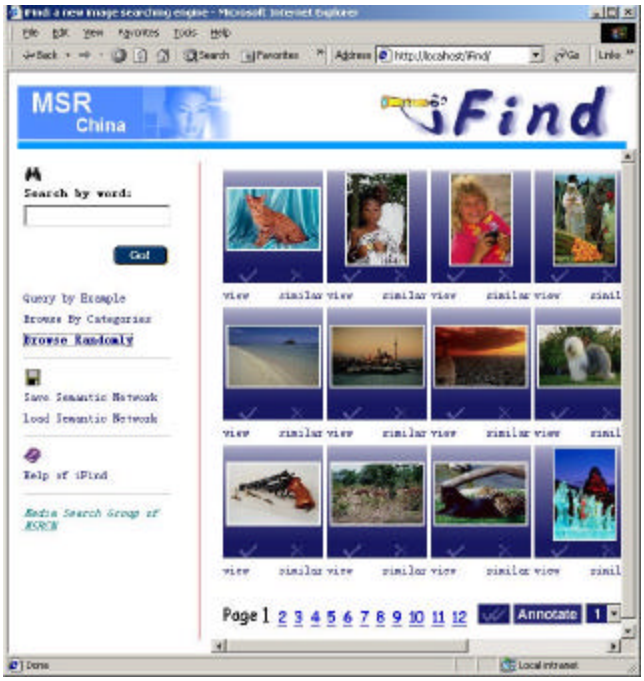

Figure 2: Main user interface.

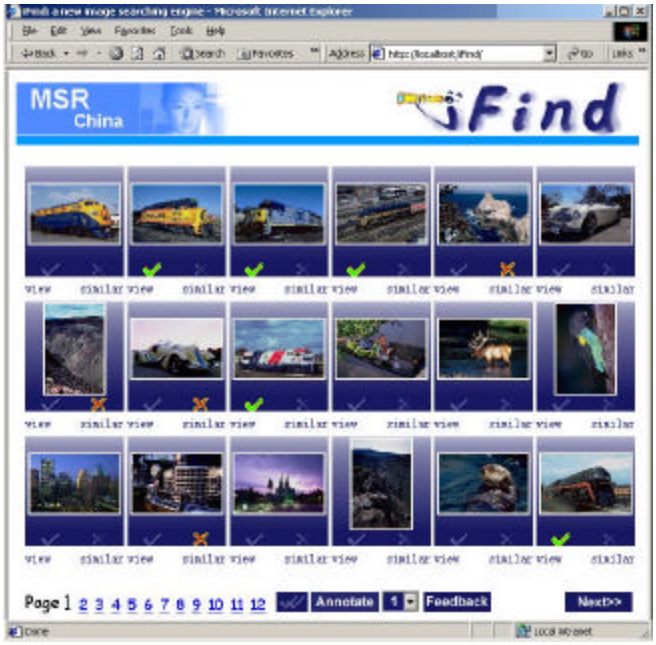

Figure 3: The query result page

The user is able to select multiple images from this page and click on the "Feedback" button to give positive and negative feedback to our system. The images with blue background indicate a positive feedback while images with a red background indicate a negative feedback. Images with gradient background are not considered in the relevance feedback process. The system presents 240 images for each query. The first 100 images are actually retrieved using the algorithm outlined in Section 3. The

next 120 images are randomly selected from each category. The final 20 images are randomly selected regardless of categories. The purpose of presenting the randomly selected images would be to give the user a new starting point if none of the images actually retrieved by our system can be considered relevant. New search results will be presented to the user as soon as the "Feedback" button is pressed. At any point during the retrieval process, the user can click on the "View" link to view a particular image in its original size, or click on the "Similar" link to perform an example based query. One point of detail to note is that if the user enters a set of query keywords that cannot be found in the semantic network, the system will simply output the images in the database one page at a time to let the user browse through and select the relevant images to feedback into the system.

4.2 Results

Here are some experimental results that we have gathered from our system to validate some simple assumptions and demonstrate its effectiveness. Because we are interested in examining how the semantic network evolves with an increasing number of user feedbacks, we select a very clean but roughly labeled image set as our starting point. The dataset that we have chosen is from the Corel Image Gallery. We have selected 12,000 images and manually classified them into 60 categories.

One assumption we have made in the design of the system is that a significant portion of the total weight of all the keyword associations with an image is concentrated on a subset of keywords that are relevant to the semantic content of the image. This relationship is shown in Figure 4 with the x axis being the number of keywords associated with the image and the y axis being the average percentage of the total weight that are assigned to relevant keywords.

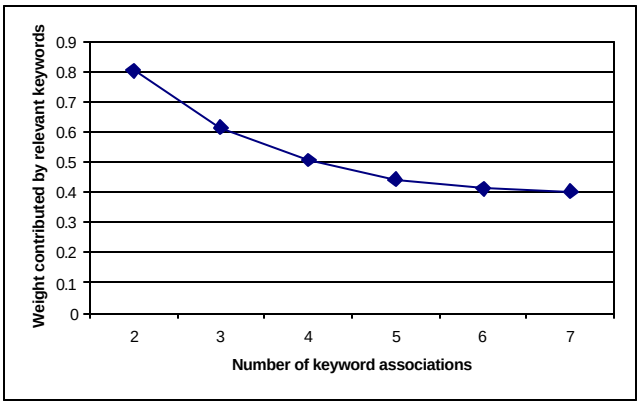

Figure 4: Keyword relevance VS keyword count.

To obtain the graph shown in Figure 4, we have asked human subjects to examine the keyword association on the images having 2 to 7 keywords associated and pick out the relevant keywords. These keyword associations are obtained from the user query using the method described in Section 3. We have also verified that the keywords with large weights are indeed the relevant keywords selected by the users. From the plot of Figure 4, we can see that as the number of keyword associations increase, the percentage of the weight contributed by the relevant keywords levels off to approximately 40%. We therefore conjecture that if we rank the keywords in descending order of their associated weight and select the top few that contribute no more than 40% of the total weight, the selected keywords will be an accurate representation of the semantic meaning of the image. The verification of this conjecture is currently on the list of our future works.

Figure 5 shows the performance of our system in terms of precision and recall. We performed eight random queries on our system. We ensured that none of the query keywords are labeled on any of the images and that there are exactly 100 images with the correct semantic content in our image database. Since we have used exactly 100 images as our ground truth for each query and that we only actually retrieve 100 images, the value of precision and recall is the same. Therefore, we have used the term "Accuracy" to refer to both in our plot.

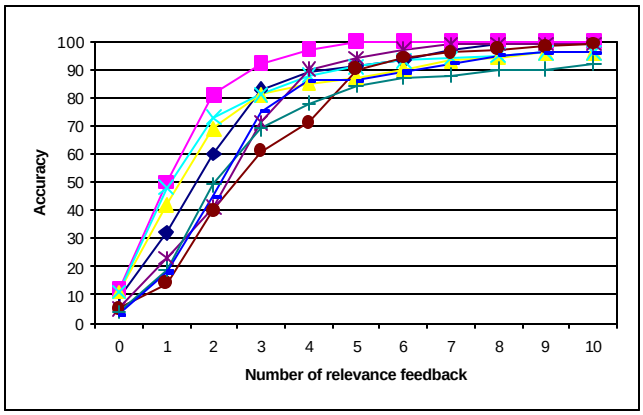

Figure 5: System performance.

As we can see from the results, our system achieves on average 80% retrieval accuracy after just 4 user feedback iterations and over 95% after 8 iterations for any given query. In addition, we can clearly see that more relevant images are being retrieved as the number of user feedbacks increase. Unlike some earlier methods where more user feedback may even lead to lower retrieval accuracy, our method proves to be more stable.

In addition to verifying the effectiveness of our system through the performance measure shown in Figure 5, we have also compared it against other state of the art image retrieval systems. We have chosen to compare our method with the retrieval technique used in the CBIR system [7]. The comparison is made through 8 sets of random queries with 10 feedback iterations for each set of query and the number of correctly retrieved images is

counted after each user feedback. The average accuracy is then plotted against the number of user feedbacks. The result is shown in Figure 6.

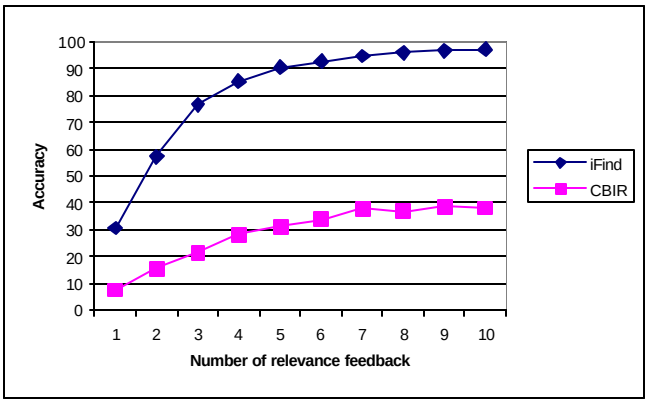

Figure 6: Performance comparison.

It is easily seen from the above result that by combining semantic level feedback with low-level feature feedback, the retrieval accuracy is improved substantially.

5. CONCLUSION

In this paper, we have presented a new framework in which semantics and low-level feature based relevance feedbacks are combined to help each other in achieving higher retrieval accuracy with lesser number of feedback iterations required from the user. The novel feature that distinguished the proposed framework from the existing feedback approaches in image database is twofold. First, it introduces a method to construct a semantic network on top of an image database and uses a simple machine learning technique to learn from user queries and feedbacks to further improve this semantic network. In addition, a scheme is introduced in which semantic and low-level feature based relevance feedback is seamlessly integrated. Experimental evaluations of the proposed framework have shown that it is effective and robust and improves the retrieval performance of CBIR systems significantly.

We have chosen to use the approach summarized in [7] as our low-level feature based feedback component. However, it can be easily demonstrated that this framework is general enough to allow any low-level feedback method to be incorporated. As a future work, we will study the possibility to incorporate the approaches proposed in [2, 4] to further improve the performance of the *iFind* system.

6. REFERENCES

[1] Buckley, C., and Salton, G. "Optimization of Relevance Feedback Weights," in Proc of SIGIR'95.

[2] Cox, I.J., Miller, M.L., Minka, T.P., Papathornas, T.V., Yianilos, P.N. "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments" IEEE Tran. On Image Processing, Volume 9, Issue 1, pp. 20-37, Jan. 2000.

[3] Ishikawa, Y., Subramanya R., and Faloutsos, C., "Mindreader: Query Databases Through Multiple Examples," In Proc. of the 24th VLDB Conference, (New York), 1998.

[4] Lee, C., Ma, W. Y., and Zhang, H. J. "Information Embedding Based on user's relevance Feedback for Image Retrieval," Technical Report HP Labs, 1998.

[5] Paek S., Sable C.L., Hatzivassiloglou V., Jaimes A.,Schiffman B.H., Chang S. F., Mckeown K.R, "Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs", SIGIR'99.

[6] Rui, Y., Huang, T. S., and Mehrotra, S. "Content-Based Image Retrieval with Relevance Feedback in MARS," in Proc. IEEE Int. Conf. on Image proc., 1997.

[7] Rui, Y., Huang, T. S. "A Novel Relevance Feedback Technique in Image Retrieval," ACM Multimedia, 1999.

[8] Salton, G., and McGill, M. J. "Introduction to Modern Information Retrieval," McGraw-Hill Book Company, 1983.

[9] Shaw, W. M. "Term-Relevance Computation and Perfect Retrieval Performance" Information processing and Management.