Semi-Automatic Image Annotation

Liu Wenyin[1], Susan Dumais[2], Yanfeng Sun[1], HongJiang Zhang[1],

Mary Czerwinski[2] and Brent Field[2]

[1]Microsoft Research China, 49 Zhichun Road, Beijing 100084, PR China

[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052

{wyliu, sdumais, yfsun, hjzhang, marycz, brentfi}@microsoft.com

Abstract: A novel approach to semi-automatically and progressively annotating images with keywords is presented. The progressive annotation process is embedded in the course of integrated keyword-based and content-based image retrieval and user feedback. When the user submits a keyword query and then provides relevance feedback, the search keywords are automatically added to the images that receive positive feedback and can then facilitate keyword-based image retrieval in the future. The coverage and quality of image annotation in such a database system is improved progressively as the cycle of search and feedback increases. The strategy of semi-automatic image annotation is better than manual annotation in terms of efficiency and better than automatic annotation in terms of accuracy. A performance study is presented which shows that high annotation coverage can be achieved with this approach, and a preliminary user study is described showing that users view annotations as important and will likely use them in image retrieval. The user study also suggested user interface enhancements needed to support relevance feedback. We believe that similar approaches could also be applied to annotating and managing other forms of multimedia objects.

Keywords: image annotation, image retrieval, relevance feedback, image database, user study, performance evaluation

1 Introduction

Labeling the semantic content of images (or generally, multimedia objects) with a set of keywords is a problem known as image (or multimedia) annotation. Annotation is used primarily for image database management, especially for image retrieval. Annotated images can usually be found using keyword-based search, while non-annotated images can be extremely difficult to find in large databases. Since the use of image-based analysis techniques (what is often called content-based image retrieval) (Flickner et al., 1995) is still not very accurate or robust, keyword-based image search is preferable and image annotation is therefore unavoidable. In addition, qualitative research by Rodden (1999) reports that the most desirable search capability for managing personal digital photographs is the ability to search based on text annotations.

Currently, most of the image database systems employ manual annotation (Gong et al., 1994). That is, users enter some descriptive keywords when the images are loaded/registered/browsed. Although manual annotation of image content is considered a "best case" in terms of accuracy, since keywords are selected based on human determination of the semantic content of images, it is a labor intensive and tedious process. In addition, manual annotation may also introduce retrieval errors due to users forgetting what descriptors they used when annotating their images after a lengthy period of time. Researchers have explored techniques for improving the process of manual annotation. Shneiderman and Kang (2000) developed a direct annotation method that focuses on labeling names of people in photos. With this method, the user can simply select a name from a manually entered name list and drag and drop it onto the image to be annotated. Although it avoids most of the typing work, it is still a manual method that involves many drag and drop operations. Moreover, there are some limitations related to the

name list, especially as the list of potential nametags becomes long. Because it requires time and effort to annotate photographs manually even with improved interfaces, users are often reluctant to do it, so automatic image annotation techniques may be desirable.

Recently researchers have used the context in which some images are embedded to automatically index images. Shen et al. (2000) use the rich textual context of web pages as potential descriptions of images on the same pages. Srihari et al. (2000) extract named entities (e.g., people, places, things) from collateral text to automatically index images. Lieberman (2000) describes a system ARIA (Agent and Retrieval Integration Agent) that integrates image retrieval and use. The system uses words in email messages in which images are embedded to index those images. When textual or usage context is available this seems like a reasonable approach, although the precision of textual context is likely not as high as manual indexing. More importantly, there are many applications such as home photo albums in which there will be minimal if any collateral text to use for automatic indexing.

Ono et al. (1996) have attempted to use image recognition techniques to automatically select appropriate descriptive keywords (within a predefined set) for each image. However, they have only tested their system with limited keywords and image models, so its generalizability to a wide range of image models and concepts is unclear. Moreover, since image recognition techniques are not completely reliable, people must still confirm or verify keywords generated automatically.

In this paper, we propose a semi-automatic strategy for semantically annotating images that combines the efficiency (speed) of automatic annotation and the accuracy (correctness) of manual annotation. The strategy is to create and refine annotations by encouraging the user to provide feedback while examining retrieval results. In textual information retrieval, relevance feedback has been shown to improve retrieval accuracy in both classic retrieval evaluations and user studies (Harman, 1992, Koenemann and Belkin, 1996), and we believe that similar techniques can be used successfully in image retrieval. Our approach employs both keyword-based information retrieval techniques (Frakes and Baeza-Yates, 1992) and content-based image retrieval techniques (Flickner et al., 1995) to automate the search process. When the user provides some feedback about the retrieved images, indicating which images are relevant or irrelevant to the query keywords, the system automatically updates the association between the keywords and each image based on this feedback. The annotation process is accomplished behind the scenes except for relevant/not relevant gestures required by the user. As the process of retrieval and feedback repeats, more and more images will be annotated through a propagation process and the annotation will become more and more accurate. The result is a set of keywords associated with each individual image in the database.

The strategy of semi-automatic image annotation is better than manual annotation in terms of efficiency and better than automatic annotation by image content understanding in terms of accuracy. As we show in our experiments, the strategy is practical and fairly easy to use, although we are still iterating the design of the user interface through user studies.

We describe the strategy in detail in Section 2 and evaluate it in Section 3. We conclude the paper in Section 4.

2 The Proposed Strategy

We call the proposed strategy a semi-automatic annotation strategy in an image database system because it depends on the user's interaction to provide an initial query and feedback and the system's capability for using these annotations as well as image features in retrieval. In this section, we first briefly present a user interface framework and scenario for image search and relevance feedback in an image database system. We then present our annotation strategy and discuss other related issues.

2.1 User Interface Framework for Image Search and Relevance Feedback

A variety of user interfaces for image retrieval and relevance feedback can be used for the proposed annotation method. Any such user interface will include three parts: the query submission interface (for either a keyword query, an image example query, or a combination of the two), the image browser, and the relevance feedback interface.

A typical user scenario is the following. When the user submits a query, the system returns search results as a ranked list of images according to their similarity to the query. Images of higher similarity are ranked higher than those of lower similarity. These retrieved images are displayed in an image browser in the order given by the ranked list. The image browser can be a scrollable window of thumbnails, a paged window (browsed page by page) of thumbnails, or some other innovative display,

such as the zoomable image browser (Combs and Bederson, 1999). The user can browse the images in the browser and use the feedback interface to submit relevance judgments. The system iteratively returns the refined retrieval results based on the user's feedback and displays the results in the browser. This process is illustrated in Figure 1.

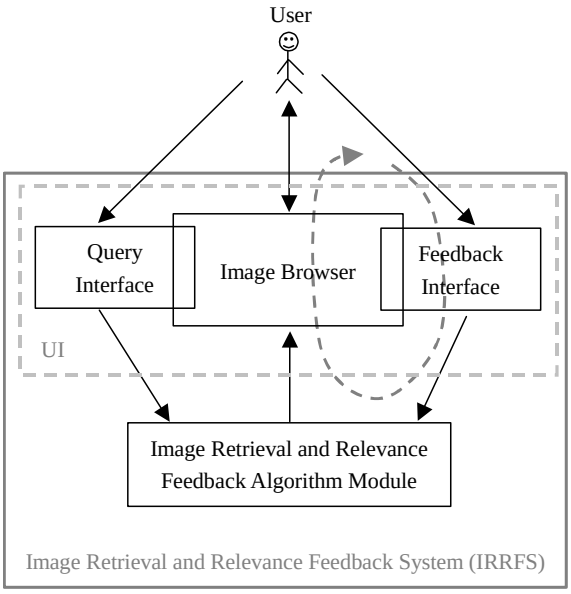

Figure 1: A typical user interface framework and a scenario of the image retrieval and relevance feedback system.

2.2 Algorithms for Matching and Search Refinement

In the image retrieval and relevance feedback mechanism, the overall similarity of an image to the query and/or feedback images can be computed using both keywords and visual features. There are many ways to combine keyword and image features in retrieval, but they are not the focus of this paper. In our prototype system, we simply use the weighted sum of the keyword-based similarity and the visual-feature-based similarity to calculate the overall score of an image. Similarity measures based on only low-level visual features are known as content-based image retrieval. The content-based image retrieval technique employed in the strategy can be any one in existence, e.g., Flickner et al. (1995), or others developed in the future. The matching could be based on any kind of visual features, e.g., color features, texture features, and shape features, using any similarity model (see Jain et al. (1995) for a survey of possible techniques). Similarly, the keyword-based similarity assessment method can be any one which is either available or may be developed in the future. We actually employed the matching method used by Lu et al. (2000).

Relevance feedback can be an effective approach to refine retrieval results by estimating an ideal query using the positive and negative examples given by the users. Each time the user provides feedback about the retrieved images, the similarity of each image in the database to the query will be recalculated according to the feedback images using some relevance feedback method. In the feedback process, any feedback method, e.g., Cox et al. (1996), Rui and Huang (2000), or Vasconcelos and Lippman (1999) can be used. The relevance feedback framework proposed by Lu et al. (2000) is preferable for our implementation because it uses both semantics (keywords) and image-based features during relevance feedback.

2.3 Semi-Automatic Annotation During Relevance Feedback

After the user provides feedback about the retrieved images, the system annotates them. The annotation process and the relevance feedback process are integrated together. Relevance feedback allows more relevant images to be shown in the top ranks and provides the user with more opportunity to see them, confirm them, and therefore annotate them through further iteration.

In our proposed approach, annotations are updated automatically whenever relevance feedback is provided, as shown in the scenario in Figure 1. Specifically, the user can submit a query consisting of one or more keywords and the system then uses the keyword(s) to automatically search in the database for those images relevant to the keyword(s). There are two situations to consider at this point. In the first case, there are no images in the system that have been annotated with the query keyword(s). In the second case, some images are already annotated with the keyword(s) that match the query. (The user would have manually annotated these images when the images were registered into the system, or the system had already progressively tagged the images with the keyword(s) through iterative relevance feedback.)

In the first case, the system only returns a random list of images since no keyword is matched and no image that is semantically relevant to the query keyword(s) can be found. Not surprisingly, this might be confusing to a user of the system, and we discuss in detail below how the user interface can be designed to address this problem. In the second case, annotated images with the same or similar keyword(s) as specified by the query are retrieved

and shown in the browser as top ranking matches. In addition, more images will be added to the browsing list: a set of images found based on their visual feature similarity to the images matched with the query (as discussed in more detail in the next subsection), and/or a set of randomly selected images.

From the retrieved image list, the user may use the relevance feedback interface to tell the system which images are relevant or irrelevant. For each of those relevant images, if the image has not been annotated with the query keyword yet, the image is annotated with the keyword with an initial weight of 1. If the image has already been annotated, the weight of this keyword for this image is increased with a given increment of 1. For each of the irrelevant images, the weight of this keyword is decreased to one fourth of its original weight. If the weight becomes very small (e.g., less than 1), the keyword is removed from the annotation of this image. The result is a set of keywords and weights associated with each individual image in the database. The links between keyword annotations and images form a semantic network. In the semantic network, each image can be annotated with a set of semantically relevant keywords and conversely, the same keyword can be associated with many images.

As we presented above, the annotation strategy is a process of updating the keyword weights in the semantic network. There may be many methods that can be used to re-weight the keywords during the annotation process. The above re-weighting scheme is simply a specific one we chose for our initial investigation of this strategy.

The relevance feedback process is repeated and both annotation coverage and annotation quality of the image database will be improved as the query-retrieval-feedback process iterates.

2.4 Possible Automatic Annotation

When one or more new (un-annotated) images are added into the database, an unconfirmed automatic annotation process can take place. The system automatically uses each new image as a query and performs a content-based image retrieval process. For the top N (which can be determined by the user) similar images to a query, the keywords in the annotations would be analyzed. A list of keywords sorted by their frequency in these N images is stored in an unconfirmed keyword list for the input (query) image. The new image is thus annotated (though virtually and without confirmation) using the unconfirmed keywords. Even unconfirmed keywords can be useful to augment retrieval techniques based solely on visual features. It is important to note that unconfirmed keywords would receive less weight than manually added keywords in the matching algorithm. An interface option could be provided to let the user manually confirm these keywords. The user could only confirm one or two keywords if he or she is reluctant to confirm all relevant keywords. The unconfirmed annotations will be refined (e.g., changing unconfirmed keywords to confirmed) through daily use of the image database in the future.

3 Implementation and Evaluation

We have implemented the proposed image annotation strategy in our *MiAlbum* system. The *MiAlbum* prototype is a system for managing a collection of family photo images, which typically have no initial annotations. The user can import images, search for images by keywords (once they have been added), and find similar images using content-based image retrieval techniques. In this system, we have implemented the keyword and content-based relevance feedback method presented by Lu et al. (2000) as our image retrieval and feedback algorithm. We augment this core matching and feedback technique with our proposed image annotation strategy as part of the user interface, which can in turn facilitate the text-based image search. In this section, we evaluate this annotation method from the perspective of both efficiency and usability.

3.1 Performance Evaluation

We first evaluated the annotation performance of the proposed approach in ideal, simulated cases. In order to objectively and comprehensively evaluate the performance of the annotation strategy, we needed to build a ground truth database, design an automatic experiment process, and define quantitative performance metrics.

The ground truth database is composed of 122 categories, each consisting of 100 images. Therefore, there are in total 12,200 images in the database, most from Corel image databases. Example categories included “people”, “horse”, and “dog”. Images within each category are considered similar or relevant to each other. We assume that each image is characterized by only one keyword, which is exactly its category name. That is, if the query keyword is the category name or the query example is an image in this category, all images in the same category are expected to be found by the image retrieval systems.

We designed an automatic experimental process to test our proposed strategy with all 12,200 images as follows. The system uses each category name as a

keyword query for image retrieval. The result will be a random list at first since we assume that there are no annotations at all in the database. The system automatically selects images from this category that appear in the first 100 retrieved images as positive feedback examples and the rest of the first 100 as negative feedback examples. These simulated relevance judgments serve as input to the first iteration of relevance feedback. If there are no relevant images in the top 100 images, all images are taken as negative examples for feedback. Using such a relevance feedback process, in which both keyword matching and content-based technique have been used, the system is able to return results with more relevant images. The top 100 images undergo the same process for the second and further iterations of feedback. As the number of iterations increases the result becomes better and better. The system repeats the feedback for 20 iterations (in our experiments) and records performance statistics at each iteration. Two common performance measures are retrieval accuracy and annotation coverage.

Annotation coverage is the percentage of annotated images in the database. We are interested in how many images can be annotated using the proposed strategy at a given iteration stage. An efficient method should need fewer iterations to get better annotation coverage.

Retrieval accuracy is how often the labeled items are correct. Since, at each iteration, the positive examples are automatically annotated with the query keyword (the category name), retrieval accuracy is the same as annotation coverage in our experiment. Since each image has 100 similar/relevant ground truth images in the database and we examine the first 100 ranked images in the retrieval list, the recall and precision metrics are exactly the same and are referred to as retrieval accuracy in our experiments. The retrieval accuracy curve of our *MiAlbum* system is shown in Figure 2. When there is no initial annotation at all, it is possible to achieve about 50% annotated images with an average of 10 iterations of relevance feedback for the 122 categories/keywords in our experiment.

We also test the retrieval accuracy of the system when there are 10% initial annotations in the database as shown in Figure 2. As we can see, the retrieval accuracy improves faster in the initial several feedback iterations than without any annotation and also asymptotes at a higher level. Consequently, the annotation strategy is more efficient when there are some initial manual annotations.

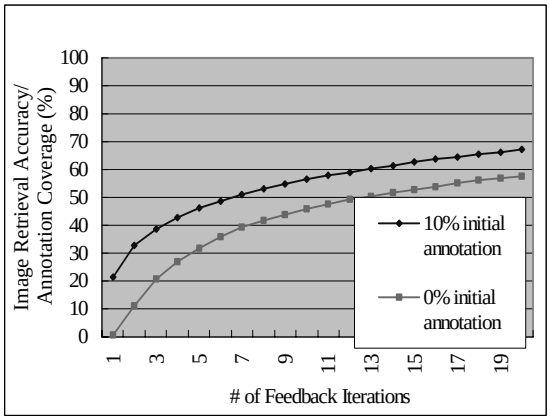

Figure 2: Image retrieval accuracy of the *MiAlbum* system and the annotation coverage of the proposed annotation strategy.

In these experiments, we found that the retrieval accuracy (or the annotation coverage) increases slowly in the first several feedback iterations for some queries (e.g., query 2 in Figure 3) compared to other queries (e.g., query 1 in Figure 3). In the slowly increasing cases, some initial manual annotation will greatly increase the annotation efficiency, since further retrieval/feedback accuracy increases very fast, as we can see from Figure 3.

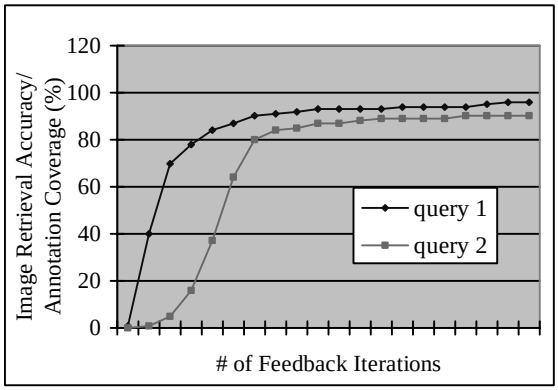

Figure 3: Image retrieval accuracy of the *MiAlbum* system of two specific queries.

3.2 Usability Studies

Because our proposed approach depends on the discoverability and ease of feedback, we performed some usability studies using the *MiAlbum* prototype implementation of the process. As part of a systematic series of studies on user interfaces for managing home photo albums, we explored several techniques for allowing users to more efficiently organize their personal photographs, including annotation, automatic clustering, and semi-automatic annotation. In this sub-section, we focus on the aspects having to do with the semi-automatic

annotation method. Participants in the studies were given a series of tasks (e.g., import pictures, annotate pictures, find pictures, use relevance feedback), and asked to think aloud as they worked on each task. There were no tutorials on any aspect of the system, so subjects had to discover all the functionality on their own. At the end of the study participants completed a short questionnaire.

Figure 4 shows a screen shot of the user interface we used for text-based search, relevance feedback, and semi-automatic annotation in our study. There are three main regions in the user interface. The upper left pane is for structured folder and category views, which are not relevant for our example. The bottom left pane is the query area, where the user can type in a keyword (in the small text box) or drag an example image (into the larger grey region). Queries can contain keywords, images, or a combination of the two with the relative importance of the two controlled by the slider. Finally, the large right pane is the image browser. Results are shown in the image browser as thumbnail images. After a search, each image contains thumbs-up and thumbs-down icons used for relevance feedback. Also, keyword annotations associated with each image are shown directly below it (e.g., 'jessica' for the first image, 'jessica' and 'dog' for the second image, etc.).

In the search scenario shown in Figure 4, a user entered the keyword query, "Jessica" in the bottom left pane, and pressed the button "Search" (not shown in this view, as will be clarified below). The retrieved images are returned along with thumbs up/down indicators for each thumbnail image in the image browser in the right-hand side Figure 4. When the user wants to provide feedback, he/she can click on the thumb up indicator for positive feedback (which means this image is relevant to the query), or click on the thumb down indicator for negative feedback (which means this image is non-relevant to the query). If some images are selected for positive or negative feedback, the "Search" button changes to "Refine". In this example, the third images in the second and third rows are given a thumb's down, and the fourth images in the second and third rows are given a thumb's up. After the user selects 'Refine',

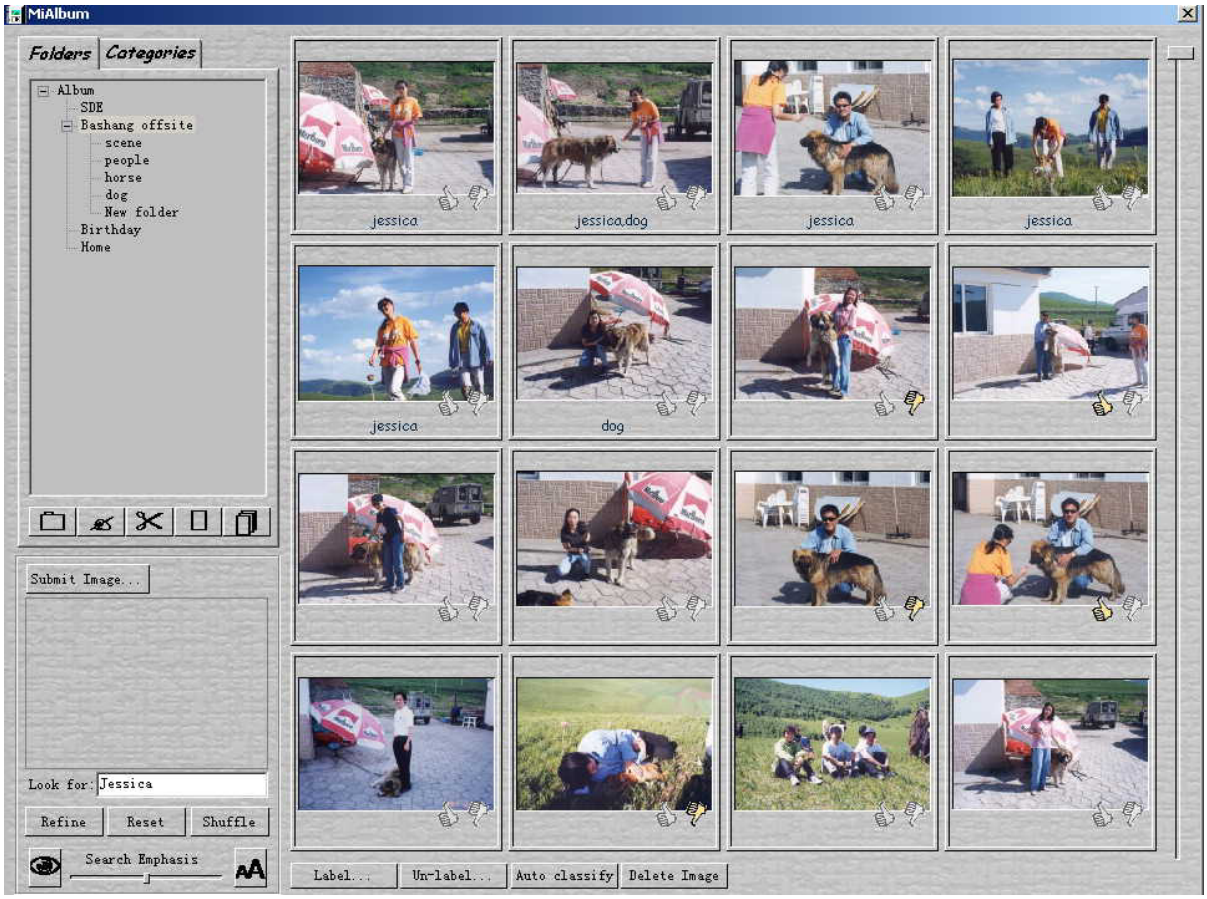

Figure 4: A screen dump of *MiAlbum*.

the search results are improved using the relevance judgments provided, and the association between the keywords and each image based on this feedback is updated. More specifically, the query keywords are added to those positive feedback images as annotations, or removed from those negative feedback images if they were previously annotated with the query keyword. The updated annotations can be used in subsequent retrieval. After using the system in this way for sometime, both the annotation coverage and the search accuracy will be greatly improved.

We next focus on the questionnaire results to gain an understanding of the overall usability of *MiAlbum* and the discoverability of search and feedback techniques. One of the highest rated questionnaire items in our user studies was the overall ease of entering annotations for images (an average of 5.6 on a 7-point scale). Participants also said that it was easier to search once photos had been annotated (an average of 6.3 on a 7-point scale), and indeed they remembered which pictures they had annotated and were faster at finding annotated versus non-annotated images. Overall ratings of the intuitiveness of refining the search to get better results (using our semi-automatic annotation approach) were about average (an average of 4.1 on a 7-point scale). There were some positive comments about the feedback and semi-automatic annotation, e.g., subjects liked: "When using the up and down hands the software automatically annotated the photos chosen", and "The ability to rate pictures on like/dislike and have the software go from there". There were also some negative comments focusing primarily on difficulties in understanding the feedback process in general, and details of exactly how the matching algorithm operated.

The results of our user studies show that we need to do additional work to improve the discoverability of relevance feedback since it is the key to the effective use of our semi-automatic annotation technique. We know that when users provide feedback, in this and other systems, the accuracy of their searches improves. However getting people to discover and use relevance feedback has been difficult, even in text retrieval systems where it was originally developed. For instance, Koenemann and Belkin (1996) have shown that increasing the transparency of relevance feedback improves how effectively users take advantage of it. But even they had issues with discoverability of relevance feedback and gave users a 45-minute tutorial about the retrieval system and relevance feedback before their experiment. In many applications tutorials are not possible, so we are looking at ways of improving the thumbs up/down metaphor and of streamlining the refinement process. In addition to improving the discoverability of feedback, we need to improve the participants' understanding of the matching process. The matching is complex since it includes both image and keyword/annotation features, but perhaps some of Koenemann and Beklin's ideas about transparency would help here. This remains as future work as the user interface is iteratively redesigned and improved.

4 Concluding Remarks

We present a semi-automatic annotation strategy that employs available image retrieval algorithms and relevance feedback user interfaces. We have used this strategy in our *MiAlbum* system and demonstrated that this strategy is effective for annotating images both in usability studies and in simulated performance evaluations.

The semi-automatic image annotation strategy can be embedded into the image database management system and is implicit to users during the daily use of the system. The semi-automatic annotation of the images will continue to improve as the usage of the image retrieval and feedback increases. It therefore avoids tedious manual annotation and the uncertainty of fully automatic annotation. This strategy is especially useful in a dynamic database system, in which new images are continuously being imported over time.

The evaluation experiments show that this strategy is very efficient compared to manual annotation and more accurate than automatic annotation. However, the performance of the annotation strategy relies heavily on the performance of content-based image retrieval (CBIR) and relevance feedback algorithms used in the framework, especially when there is no initial annotation in the database at all. For those queries resulting in low CBIR performance, some initial annotation (including manual annotation) can help increase the annotation efficiency. CBIR and relevance feedback together allow more relevant images to be shown in the top ranks of the retrieval results and provides the user with more opportunity to see and confirm relevant items through further iteration. The annotation efficiency is therefore improved.

Preliminary usability results are promising, but further user interface refinements will be needed to

improve the discoverability of the feedback process and the underlying matching algorithm.

As content-based retrieval techniques of multimedia objects become more effective, we believe a similar semi-automatic annotation framework can be used for other multimedia database applications.

References

Combs TTA and Bederson BB (1999) Does zooming improve image browsing. In: *Proceedings of ACM Digital Libraries (DL99)*, pp. 130-137.

Cox IJ, Miller ML, Omohundro SM, and Yianilos PN (1996) PicHunter: Bayesian relevance feedback for image retrieval. In: *Proceedings of ICPR96*, pp. 361-369.

Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, and Yanker P (1995) Query by image and video content: The QBIC system. *IEEE Computer*, 28(9), 23-32.

Frakes W and Baeza-Yates R (1992) (eds.) *Information Retrieval: Data Structures and Algorithms*. Prentice Hall.

Gong Y, Zhang H, Chuan HC, and Sakauchi M (1994) An image database system with content capturing and fast image indexing abilities. In: *Proceedings of IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 121-130.

Harman D (1992) Relevance feedback and other query modification techniques. In: Frakes W and Baeza-Yates R (Eds.), *Information Retrieval: Data Structures and Algorithms*, Chapter 11, pp. 241-263. Prentice Hall.

Jain R, Murthy SNJ, and Chen PLJ (1995) Similarity measures for Image Databases. In: *Proceedings of IEEE Conf. On Fuzzy Systems*, vol. 3, pp. 1247-1254.

Koenemann J and Belkin N (1996) A case for interaction: A study of interactive information retrieval behavior and effectiveness. In: *Proceedings of CHI'96*, pp. 205-212.

Lieberman H (2000) An agent for integrated annotation and retrieval of images. Paper presented at Workshop on *Personal Photo Libraries: Innovative Designs.* University of Maryland, June 1, 2000.

Lu Y et al. (2000) A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: *Proceedings of ACM Multimedia2000*, pp. 31-38.

Ono A, Amano M, Hakaridani M, Satou T, and Sakauchi M (1996) A flexible content-based image retrieval system with combined scene description keyword. In: *Proceedings of IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 201-208.

Rodden K (1999) How do people organise their photographs? In: *BCS IRSG 21st Ann. Colloq. on Info. Retrieval Research*.

Rui Y and Huang TS (2000) Optimizing learning in image retrieval. In: *Proceedings of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR2000)*, pp. 236-244, Hilton Head, SC.

Shneiderman B and Kang H (2000) Direct annotation: A drag-and-drop strategy for labeling photos. In: *Proc. International Conference Information Visualisation (IV2000)*, pp. 88-98. London, England.

Shen HT, Ooi BC, and Tan KL (2000) Giving meanings to WWW images. In: *Proceedings of ACM Multimedia2000*, pp 39-48.

Srihari RK, Zhang Z, and Rao A (2000) Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2, 245-275.

Vasconcelos N and Lippman A (1999) Learning from user feedback in image retrieval system. In: *Proceedings of NIPS'99*, pp. 843-849 Denver, Colorado.