This paper is published in the Proc. of 2002 IEEE Int'l Conf. on Multimedia and Expo, Lausanne, Switzerland, Aug, 2002.

IMAGE RETRIEVAL AND RELEVANCE FEEDBACK USING PEER INDEXING

Jun Yang [1,2] *Qing Li*[1] *Yueting Zhuang*[2]

[1] City University of Hong Kong, Hong Kong SAR, China, {itjyang, itqli}@cityu.edu.hk

[2] Zhejiang University, Hangzhou, China, yzhuang@cs.zju.edu.cn

ABSTRACT

We present the idea of peer indexing—indexing an image by semantically correlated images—and its application in image retrieval. A learning strategy is suggested for automatic acquisition of peer indices from user feedbacks, and the similarity metric for peer index is formulated. A cooperative framework is proposed under which peer index is integrated with low-level features for image retrieval and relevance feedback. Encouraging results on both short-term and long-term retrieval performance of our approach are shown by experiments.

1. INTRODUCTION

Image indexing techniques are the foundation of image retrieval and therefore have a great impact on the retrieval performance. The popular indexing techniques can be roughly classified into three categories—indexing by keywords, low-level features, and visual primitives. Keyword index is a list of descriptive keywords attached with each image as its annotation, based on which images can be searched by keywords with high accuracy. Despite its effectiveness for retrieval purpose, keyword index cannot be generated automatically and is usually annotated manually. Low-level features, such as color histogram and texture, are directly computable from images and widely used in content-based image retrieval (CBIR) [4]. However, since low-level features are unable to capture precisely the high-level semantics of images, the performance of CBIR is unsatisfactory. In the third indexing scheme, each image is segmented into blocks or visual objects, which are clustered in the feature space, with each cluster defined as a visual primitive [2] [5]. Therefore, images can be indexed and searched by visual primitives. Due to its reliance on low-level features, this indexing method still suffers from the problem of the semantics-feature gap.

In this paper, we present a novel idea on image indexing—*peer indexing*—and its application in image retrieval. The peer index of an image is a list of semantically correlated images termed as *visual keywords*. Due to its analogy with keywords, peer index is weighted and its mutual similarity is calculated using mature techniques of text-based information retrieval (IR). A learning strategy is suggested to derive peer indices progressively from user-provided feedbacks. Furthermore, we propose a *cooperative framework* under which peer indices are integrated with low-level features for image retrieval and relevance feedback. Experiments conducted on real-world images manifest that our approach can achieve retrieval performance much higher than that of CBIR systems. On the other hand, peer index serves as the "memory" of historical feedback information, which also helps improve the long-term retrieval performance.

The rest of this paper is organized as follows. In Section 2, we describe the representation, acquisition strategy, and similarity metric of peer index. The cooperative framework for image retrieval and relevance feedback is introduced in Section 3. Section 4 presents a comparison of our work with related works. Performance evaluation is presented in Section 5 and conclusion is given in Section 6.

2. PEER INDEX: REPRESENTATION, ACQUISITION, AND SIMILARITY METRIC

Peer indexing is based on a simple and intuitive notion: each image embodies an implicit semantic concept, which becomes emergent through its correlation with other images. This notion is analogous to the idea of estimating the impact factor of a scientific journal based on the citations of its papers by the papers of other journals, or calculating the degree of "authority" of a web page through its hyperlinks with other web pages.

2.1. Index Representation

In peer indexing, each image plays a dual role—an image to be indexed and a visual keyword used to index other images. We name it as visual keyword based on the notion that an image is a visual representation of a semantic concept. The peer index of an image I_i is represented as a list of weighted visual keywords:

$$P(I_i) = \{< vk_{i_1}, w_{i_1} >; \cdots; < vk_{i_k}, w_{i_k} >; \cdots; < vk_{i_N}, w_{i_N} >\} \qquad (1)$$

where vk_{i_k} is a visual keyword corresponding to an image semantically correlated with I_i, with the weight w_{i_k} indicating the strength of the correlation.

2.2. Learning Strategy for Index Acquisition

To avoid building all the peer indices manually, we suggest a simple learning strategy to derive them from the statistics of user feedback information. This strategy is embedded in the process of relevance feedback. When a user submits a sample image as the initial query and evaluates some of the retrieved images as relevant or irrelevant, every relevant image and the sample image are added into each other's peer index, while every irrelevant image and the sample image are removed from each other's peer index (if exists). The algorithm describing the learning strategy is presented as follows:

1. *Collect the sample image I_S, the set of relevant examples I_R, and the set of irrelevant examples I_N.*
2. *For each $I_i \in I_R$, if I_i does not exist in $P(I_S)$, add it as a visual keyword into $P(I_S)$ with the initial weight set to 1. Otherwise, increase the weight of I_i in $P(I_S)$ with an increment of 1. Similarly, I_S is also added into $P(I_i)$ or has its weight in $P(I_i)$ increased.*
3. *For each $I_i \in I_R$, if I_i exists in $P(I_S)$, divide its weight by a factor of 5. If the resulting weight is below 1, remove I_i from $P(I_S)$. Similarly, I_S is removed from $P(I_i)$ or has its weight in $P(I_i)$ decreased.*

As more user feedbacks are conducted, the peer indices of images are improved both in coverage and in quality (reflected by weights). An advantage of this strategy is that, instead of manual

indexing, it exploits the interactions of the entire population of users to derive peer indices in an automatic and progressive manner. Therefore, not only is the great manual effort relieved, but also the subjective errors are reduced which are more likely to be made by one or a few human indexers.

2.3. Semantic Metric

Because of the analogy between real keywords and peer index as a list of visual keywords, many mature techniques developed for text-based information retrieval (IR) can be applied to peer index as well. Among them, term weighting is a technique of assigning weights on the keywords according to their importance in a document. A well-known term weighting scheme is the so-called TF*IDF, which considers two factors: (1) *term frequency* (TF) as the frequency of a keyword in the document, and (2) *inverse document frequency* (IDF), which captures the discriminative power of a keyword by considering the number of documents in which it appears. In peer index, the weight w_{i_k} attached with each visual keyword vk_{i_k} only corresponds to the first factor, and thus it needs to be adjusted to include the discriminability factor:

$$w'_{i_k} = w_{i_k}\left(\log\frac{M}{M_{i_k}} + 1\right) \tag{2}$$

where M is the total number of images, and M_{i_k} is the number of images whose peer index has vk_{i_k} in it. Thus, a visual keyword concentrating on a few images is weighted higher than the one spreading over many images, since widely distributed visual keywords are less capable of differentiating among images.

The similarity between the peer indices of two images is calculated by *cosine similarity,* a metric extensively used in text-based IR field. That is, we treat the peer index of an image as a vector, with each visual keyword corresponding to a dimension, and its weight (after refinement by Eq.2) as the length on that dimension. Thus, the similarity between two peer indices can be transformed into cosine value of the angle formed by their corresponding vectors, formulated as:

$$R_{ij} = \frac{P(I_i)\cdot P(I_j)}{\|P(I_i)\|\|P(I_j)\|} \tag{3}$$

where $\|\ \|$ is the norm of a vector, and $\cdot$ is dot product.

3. A COOPERATIVE FRAMEWORK FOR IMAGE RETRIEVAL AND RELEVANCE FEEDBACK

There is no doubt that images matched by peer index are highly relevant to the sample image, since peer index contains the user-perceived semantic correlations between images. Nevertheless, an image retrieval system purely based on peer index is rather limited in that it can only explore the images with non-empty peer index. Contradictorily, according to Section 2.2, an image has to be retrieved and labeled as a feedback example in order to obtain its peer index. On the other hand, even if the peer index is available for all the images in the database, no images can be matched for a query formed by a new image without peer index.

To reach its maximum potential, peer index should not be used only as the memory of previously retrieved relevant images, but as semantic clues through which more relevant images that have not been retrieved before can be discovered. For this purpose, we integrate low-level features, which describe the visual aspect of images, as the complement of peer index, which models the semantic correlations between images. A cooperative framework is proposed under which these two types of index can work together for image retrieval and relevance feedback.

3.1. Two-Pass Image Retrieval

The specific low-level features used in our approach include 256-d HSV color histogram, 64-d Lab color coherence, and 32-d Tamura directionality. Generalized Euclidean distance is used as the similarity metric, in which the importance of various features is reflected by weights. In our approach, the query is formulated as a sample image, which can be either a new image submitted by the user or an existing one selected from the database. The low-level features of the sample are called *query vector*, because all the features can be represented as vectors in the feature space.

The retrieval is conducted in a two-pass process, with each pass focusing on one type of index. In the first pass, the similarity of each candidate image (i.e., images in the database) with the sample image is calculated based on peer index using Eq.3. The second pass can be described as a "pseudo" feedback process, in which the images matched in the first pass (those with non-zero similarity) and their similarity score are fed into a low-level feature based relevance feedback algorithm [3] to optimize the query vector and the weights of different features in the Euclidean similarity function. Finally, the overall similarity S_i* of image I_i with respect to the sample image is calculated by the following equation:

$$S_i^* = (1 + R_i)S_i \tag{4}$$

where R_i is the peer index similarity of I_i with the sample image, which is obtained in the first pass, and S_i is I_i's low-level feature similarity with the query vector calculated by the Euclidean similarity function, both of which have been optimized in the second pass. As an exception, if no images are matched in the first pass, the feedback process in the second pass cannot be conducted. In this case, we skip the second pass and directly calculate the image similarity by Eq.4 using the initial query vector and the Euclidean similarity function.

The particular relevance feedback algorithm used in our approach is the one proposed by Rui et al. in [3]. The choice is based on two considerations: (1) It employs a sophisticated and generic model, which can be reduced to many other feature-based feedback methods as its special cases. (2) The model accommodates multiple types of low-level features, which is the case in our approach. The input to this algorithm is a list of example images that are relevant to the query, each attached with a score indicating the degree of the relevance. Based on the input, it formulates a global optimization problem, the solution to which is the optimal query vector and the optimal weights for each feature in the Euclidean similarity function.

3.2. Relevance Feedback

To refine the retrieval results, the user can label the retrieved images as relevant or irrelevant examples and activate the feedback process. Firstly, these feedback examples are passed to the learning algorithm in Section 2.2, which updates the their and the sample image's peer indices. After that, we go through the same two-pass process described in Section 3.1, which optimizes the query vector and the Euclidean similarity function. Finally, we recalculate the similarity of each image using a comprehensive similarity measure that considers both the relevant and irrelevant examples, given as:

$$S_i^* = (1+R_i)S_i + \frac{\beta}{N_R}\sum_{k\in N_R}[(1+R_{ik})S_{ik}] - \frac{\gamma}{N_N}\sum_{k\in N_N}[(1+R_{ik})S_{ik}] \quad (5)$$

where S_i^* is the overall similarity of image I_i. Similar to Eq.4, R_i is the peer index similarity of I_i with the sample image calculated by Eq.3, and S_i is I_i's low-level feature similarity with the optimized query vector. N_R and N_N are the number of relevant and irrelevant examples respectively. R_{ik} is the peer index similarity of I_i with the kth relevant (or irrelevant) example, and S_i is their low-level feature similarity calculated using the optimized Euclidean similarity function. β and γ are parameters adjusting the impact of relevant and irrelevant examples to the overall similarity function. Please note that if there is no feedback example (i.e, both N_R and N_N are zero), Eq.5 is reduced to Eq.4. Therefore, we can use Eq.5 as a uniform similarity function even for the first round of retrieval. The whole retrieval process is summarized in Fig.1.

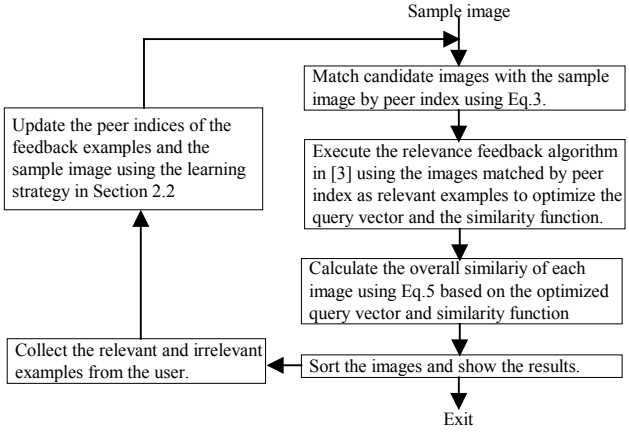

Fig.1: Cooperative framework for image retrieval and feedback

4. COMPARISON WITH RELATED WORKS

In this section, we compare our work with image retrieval approaches based on different indexing techniques, including indexing by keywords, low-level features, and visual primitives.

Keywords are effective in capturing the high-level semantics of images, and therefore keyword-based image retrieval is of high accuracy if the image annotations are available. Unfortunately, automatic image annotation is not feasible given the start-of-the-art computer vision and image understanding technologies. In most cases, keywords are assigned manually, which is not only time-consuming but also vulnerable to subjective errors. In the *iFind* system [1], a semi-automatic image annotation strategy is devised to learn the keywords from the user feedbacks. Our learning strategy suggested in Section 2.2 can be regarded as the counterpart of that annotation strategy on the peer index.

The basic paradigm of CBIR is to search for images that are visually similar to a sample image based on low-level features. A comprehensive survey of CBIR systems is given in [4]. As illustrated on a 2-D feature space in Fig.2(a), the principle of CBIR is to formulate a subspace (normally, an ellipsoid) centered at the query vector in the feature space and return all the images within it. The shape of the subspace is decided by the similarity metric, while its size depends on the number of images to retrieve (window size). As shown in Fig.2(b), the CBIR feedback algorithm re-positions (by adjusting the query vector) and deforms the subspace (by optimizing the similarity function) based on the relevant/irrelevant examples, in order to include as many relevant images as possible at a fixed window size. However, since the relevant images are defined from a semantic perspective, their distribution in the feature space can be very sparse or irregular and in this case the number of relevant images covered by even the optimal subspace is limited. Therefore, the retrieval performance of a certain query is restricted by an "upper bound" defined by the distribution of relevant images. The advantage of our approach is that, though it also relies on low-level features, it can surpass the upper bound by the peer index, which models the semantic correlations between images. As illustrated in Fig.2(c), the peer index similarity serves as the "semantic bridges" in the feature space, through which more relevant images can be reached from the sample.

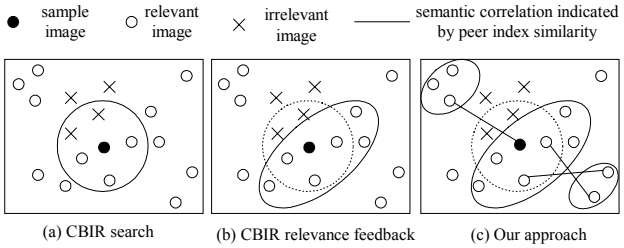

Fig.2: Comparison of our approach with CBIR approach

The representatives of indexing by visual primitives are the approaches proposed by Zhu et al. [5] and that of Paek et al. [2]. In [5], images are partitioned into fixed-size blocks, which are clustered in the feature space with each cluster defined as a *keyblock*, a visual analogy of keyword. Then an image can be encoded as a matrix of keyblocks, based on which images are searched as text documents using the text-based IR techniques. The method in [2] is very similar to the keyblock approach, except that it also incorporates the similarity of the text accompanying each image for the classification of photographs. In fact, the spirit of our approach is consistent with that of the above two approaches—indexing and searching images based on a keyword equivalent, which is visual keyword in our approach and image blocks in the keyblock approach.

5. PERFORMANCE EVALUATION

The performance evaluation is conducted on a collection 5,000 images from Corel Image Gallery, which are classified into 50 topical categories with exactly 100 images in each category. The classification is regarded as the *ground truth*, i.e., images from the same category are regarded as relevant to each other. The image categories range from "Firework" and "Iceberg", which have visually similar images, to "Insect" and "Architecture", which has visually heterogeneous images.

In our experiments, a query is formulated by randomly selecting a sample image from the test data. For each query, the system returns 100 images as the results, among which the first 90 are the images ranked top by our retrieval approach, and the last 10 images are randomly picked from the database. The random images are presented to give the user a new starting point, in case no image returned by the retrieval approach is relevant. User feedbacks are automatically generated among the 100 retrieved images according to the ground truth, i.e., images belonging to the same category as the sample are labeled as relevant examples and the rest as irrelevant ones, based on which the retrieval results are

refined using our approach. Since the number of retrieved images is equal to the number of relevant images, the value of precision and recall are the same and we use "retrieval accuracy" to refer to both of them. Note that here the highest accuracy is around 90% due to the 10 random images.

We generated totally 200 random queries (4 queries for each category) and conducted 15 rounds of feedback for each query using our approach. The peer indices of all the images are cleared before each query is processed. The average retrieval accuracy achieved at each round of feedback is shown in Fig.3. For comparison purpose, we run the same 200 queries using the CBIR relevance feedback method in [3], which is a component of our approach, and plot its performance together with that of our approach. As we can see, two approaches are at the same performance level initially (16.4%), because without initial peer index our approach is reduced to the CBIR approach. As the feedback proceeds, our approach significantly outperforms the CBIR approach, achieving accuracy as high as 84.1% after 15 rounds of feedback. (Note that the maximum accuracy is around 90%.) In comparison, the accuracy of the CBIR approach hovers around 36% after 6 feedbacks, which can be well explained by the "upper bound" theory discussed in Section 4.

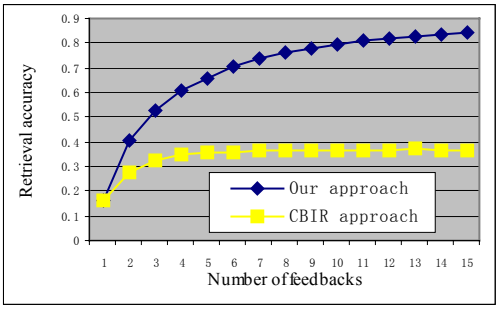

Fig.3: Performance comparison

To test the robustness of the two approaches, we also computed the standard deviation of the final accuracy (after 15 feedbacks) of different image categories, which is 10.6% for our approach versus 15.7% for CBIR approach. Looking into the final accuracy on each category, we find that the performance of the CBIR approach fluctuates greatly across categories, reaching 83.7% for a category with visually uniform images, and only 19% for another category with visually dissimilar images. In comparison, our approach performs more steadily across various categories, with the lowest accuracy 62.3%. This is a strong proof to the conclusion that our approach overcomes to a certain extent the limitation of CBIR approaches.

The previous experiment examines the performance within a single retrieval session, with each session defined as a query and the subsequent feedbacks. Besides that, we also studied the retrieval performance across different sessions, i.e., long-term performance. The experiment is designed as follows: For each category, we applied a succession of retrieval sessions, with each session consisting of a random query followed by a single round of feedback. Since the feedback in each session causes the peer indices updated, the retrieval accuracy of subsequent sessions can benefit from exploiting the peer indices learned in previous sessions. We conduct this experiment on all the categories in a random order and show the change of average accuracy in Tab.1. As we can see, the accuracy improves substantially over sessions, reaching 50% after 18 sessions. Given that only a single round of feedback is conducted in each session, our approach is very effective in promoting long-term performance.

Tab.1: Retrieval accuracy over sessions

Sessions	1	2	3	4	5	6
Accuracy(%)	16.7	23.0	26.8	29.5	32.1	34.3
Sessions	7	8	9	10	11	12
Accuracy(%)	36.0	37.6	39.2	40.9	42.6	43.4
Sessions	13	14	15	16	17	18
Accuracy(%)	44.5	45.8	47.0	48.7	49.5	50.3

6. CONCLUDING REMARKS

This paper has presented a novel indexing technique for image retrieval—peer indexing—which describes an image through a set of semantically correlated images. The methods for the acquisition, weighting, and similarity calculation of peer index are proposed. We also suggested a cooperative framework under which peer index is integrated with low-level features for image retrieval and relevance feedback. Substantial improvement over CBIR approaches on retrieval accuracy has been achieved using our approach, which also promotes the long-term performance. Our future work is to investigate the efficiency issue of peer indices, which may expand exponentially with the number of images. Efficient strategy for the storage, retrieval, and filtering of peer index will be developed to improve the efficiency.

ACKNOWLEDGMENTS

The work described in this paper was supported, substantially, by a grant from CityU (Project No. 7001073), partially by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CityU 1119/99E], and partially by a grant from the Doctorate Research Foundation of the State Education Commission of China.

REFERENCE

[1] Y. Lu, et al. "A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems," *Proc. of ACM Multimedia*, pp. 31-38, 2000.

[2] S. Paek, et al. "Integration of visual and text-based approaches for the content labeling and classification of photographs." *Proc. ACM SIGIR Workshop on Multimedia Indexing and Retrieval*, 1999.

[3] Y. Rui and T.S. Huang, "Optimizing Learning in Image Retrieval," *Proc. of IEEE CVPR*, pp. 236-243, 2000.

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years." *IEEE Trans. on PAMI,* 22(12): 1349–1379, 2000.

[5] L. Zhu, A.D. Zhang, A. Rao, R.K. Srihari, "Keyblock: an approach for content-based image retrieval." *Proc. ACM Multimedia*, pp. 157-166, 2000.