User-Powered "Content-Free" Approach to Image Retrieval

Takeo Kanade and Shingo Uchihashi

October 2004
CMU-CS-04-169

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Keywords: Image retrieval, collaborative computing, information filtering

Abstract

Consider a stereotypical image-retrieval problem; a user submits a set of query images to a system and through repeated interactions during which the system presents its current choices and the user gives his/her preferences to them, the choices are narrowed to the image(s) that satisfies the user. The problem obviously must deal with image content, i.e., interpretation and preference. For this purpose, conventional so-called content-based image retrieval (CBIR) approach uses image-processing and computer-vision techniques, and tries to understand the image content. Such attempts have produced good but limited success, mainly because image interpretation is a highly complicated perceptive process. We propose a new approach to this problem from a totally different angle. It attempts to exploit the human's perceptual capabilities and certain common, if not identical, tendencies that must exist among people's interpretation and preference of images. Instead of processing images, the system simply accumulates records of user feedback and recycles them in the form of collaborative filtering, just like a purchase recommendation system such as Amazo.com.To emphasize the point that it does not deal with image pixel information, we dub the approach by a term "content-free" image retrieval (CFIR). We discuss various issues of image retrieval, argue for the idea of CFIR, and present results of preliminary experiment. The results indicate that the performance of CFIR improves with the number of accumulated feedbacks, outperforming a basic but typical conventional CBIR system.

1 Introduction

A picture is said to be worth a thousand words. If this statement is true, it is no wonder that computerized image retrieval is a challenging task. A key to a capable image retrieval system is how to extract and describe the image contents. Contents may be described either verbally or non-verbally. Verbal descriptions, such as keywords, are suitable for human perceptions. And if obtained, keyword-based retrieval is relatively straightforward [3]. Attaching keywords to images is, however, hard; manual labeling is too expensive and automatic methods, for the moment, are not reliable.

Non-verbal descriptions used so far are computer-centric. Various image features are proposed, including color, shape, and texture. The current content-based image retrieval (CBIR) approach uses these image features to define the model of similarity or visual resemblance between images [14]. This relatively simple scheme has shown good success in various image database applications [7][11]. Some limited object recognition, such as faces, has also been used [8][12]. Further, the relevance feedback technique has produced better results by incorporating user feedback for tuning the underlying model with context-dependent variations [4][13].

Capabilities of image retrieval systems based on the above current "content-based" approachesare still severely limited. Although image features capture image characteristics, they are not always directly related to the meaning of images or image interpretation. There is a difference between what image features can distinguish and what people perceive from the image. This difference, or the "semantic gap," is the core of the limitation.

A technique called relevance feedback, where the user provides his preference and the system adjust the model, brought users into the decision making, but the advantage of having humans in the loop has not been fully exploited because the feedback was restricted to the way that predefined image features are used. The semantic gap has persisted.

We propose a new approach to image retrieval that uses user feedbacks in the form of interpretation rather than through image features, thus directly exploiting human perceptive power. We adopt collaborative filtering techniques to accumulate feedbacks of all users and use them to help future users. By bypassing image features, the performance improvement will not be restricted by the predefined capabilities of feature selection or object recognition. Our proposed system may be similar to a purchase recommendation system, like the Amazon.com, that recommends books when a user purchases a few books based on purchase history of others who have bought the same books. Observing images that a user selects in the early part of the session, it retrieves "related" images based on the accumulated usage history of him/herself *and* others.

We will name our approach "content-free" image retrieval (CFIR) in order to illustrate the point that it does not analyze image pixels. Naturally, the traditional "content-based" approach must be combined in the final system, but we will explore and emphasize the "content-free" aspect throughout this paper.

2 Image retrieval: current approaches

Image retrieval has been an active research area for the last decade [14]. In this section, we review the strength and weakness of current approaches to identify where we should focus.

2.1 Keyword-based approach

One obvious approach of image retrieval is to describe contents of images in a database verbally, typically by keywords, and apply text search techniques. The difficulty with this approach lies in how to get such text data. As manual labeling is too costly, alternative sources are necessary. Images on web pages tend to have associated text surrounding them. Also, image file names or path names may provide good descriptions of image contents [15]. Recently Ahn *et al* [1] has proposed an interesting approach to combine manual image labeling and network games, in which game participants are led to willingly do the labeling task.

Automatic classification of images or recognition of objects in them is ideal. Some success is reported using relatively simple image features, such as color and texture distributions, and certain image classes, such as indoor, outdoor, sunset, and landscape, are identified [16][17]. Lipson *et al* enhanced this scheme by introducing scene configuration [9]. Only a few objects, such as faces and cars, can be recognized reliably from general images [8][12]. Several methods have been proposed to automatically learn the relationship between image regions of specific color or pattern, and keywords [2][19]. We expect constant progress in these areas, but considering the complexity of the problem and the number of objects that we have to deal with, it will be some time before the performance of automatic image understanding becomes comparable to that of human beings.

2.2 Content-based approach

To avoid the difficulty of obtaining real contents, computer-centric image representations have been used based on image features, such as color, shape, and texture [7][11][18] (see the historical summary in [14] and [20]). Images are characterized mostly by statistical properties, such as a histogram, of those features. *Similarity* measure between images is defined and used to retrieve target images. This approach is historically called "content-based," even though the name is a somewhat inflated one.

Finding a good set of features is very critical since the rest is built upon it. Various features and associated similarity measures have been proposed to imitate human visual perception. These attempts achieved only limited success so far because human perception of images is complex and seems to be dependent on context, purpose, and individual cases.

2.3 Relevance feedback

Content-based image retrieval can be enhanced by incorporating user feedbacks into the system. Typically, as the system shows the retrieved images to the user, he/she tells the system which images in the output are more relevant or less relevant to his/her query. Given feedbacks from a user, the system determines which image features are to be used to duplicate the user's decision and make changes to the parameters or weights in the underlying model of image similarity. The feedback procedures are repeated as necessary. Many researchers have reported that improved results are obtained [4][13].

3 User-powered content-free image retrieval

Relevance feedback methods have proven that humans can play an important role in the success of image retrieval; even simple user feedbacks help improve the performance of content-based image retrieval methods. The fundamental reason for this is that people know what they are doing. Humans can immediately judge whether presented images are relevant to what he/she is looking for, although people usually may not be able to provide its complete descriptions in advance.

3.1 From "content-based" to "content-free"

We observe two different types of limitation in the way that the current content-based methods use user feedback. Firstly, because our understanding of human vision is limited, we probably do not have a correct set of image features to begin with. Therefore, perception models based on those features will not satisfy all the requirements demanded by the user feedbacks. Secondly, selecting several images several times at each session will not provide enough data to train a complex vision model. To properly adjust the underlying model with sufficient complexity requires that a large number of image samples be provided by the user.

It should be noted that it is the user who analyzes image contents, and that the feedback is the result of that analysis. Image features and similarity measures are nothing but the representation tools that aggregate his analyses into a decision making process that the "content-based" image retrieval happens to use.

A natural solution to overcome these difficulties is to bypass the image features and use the human's perceptual decisions themselves (i.e., which images are similar to which) as the representations which need to be aggregated and from which the system learns the contents of the images.

We believe that an effective image retrieval system can be realized using only the usage history of users. Imagine a user who is engaged with a system for image retrieval sessions. In each session, a user is telling the system which images are relevant or similar and which images are not. Note, however, the user is not telling why or in what sense; he/she is simply telling the decision

Now we record all of these feedbacks from all of the users. The accumulated feedbacks should work as asynchronous voting on relationships among images in the database. Once enough feedbacks are accumulated, the system can learn and summarize those relationships in a certain form. Subsequently the system retrieves relevant images for a new query from a new user using the learned relationships, and the result is expected to agree with the majority's perception. Unlike the content-based approach, this scheme lets all image processing and perception tasks be done by a population of users, and uses the learned relationships from them to do the retrieval task. Hence the name: "content-free" approach.

3.2 Collaborative filtering

The tool to accumulate user feedbacks and retrieve images for a new query is collaborative filtering. Collaborative filtering is a technique to predict preferences of one person from preferences of others [10]. Amazon.com Book Store [22] is one of the best known examples. It basically works as follows. When a customer purchases a set of books, the system looks up purchase histories of other customers who have purchased the same set of books, identifies the most popular books among those customers, and recommends the identified books. Naturally, it is likely there are only few or even no previous users who purchased the *exact* same set. Collaborative filtering techniques allow for reasoning *related* books from many samples.

An image retrieval system with collaborative filtering would work similarly. When a user forms a query by selecting a set of images from the database, the system uses usage histories of previous users (including his/her own interaction histories up to then), identifies related images that would be most frequently selected with the query images, and displays the identified images.

3.3 Premises

A few assumptions have to be satisfied for the above idea to work well with image retrieval problems. Firstly, one users relevance judgment under the same context will remain relatively stable over time. Secondly, given the same pair of images and the same context, the relevance judgments of different people are similar.

Thirdly, for the finite set of images, the number of interpretations derived from the set does not grow too fast.

The first assumption says that a user's judgment is time-shift invariant. Note that this invariance does not have to hold strictly, but it should hold to the extent that recorded feedbacks of a user will be helpful for him/her in the future.

With the second assumption, feedbacks from different users are transferable, and can be treated as an ensemble. This assumption is the basis for our collaborative solution. This does not mean, however, that the interpretation or preference of a particular image must be the same over people. Rather it requires that their distribution is similar. Although there are differences in personal preferences, we assume that people share common perceptive attributes.

While the first two are related to reusability of feedbacks, the third assumption concerns the sufficiency of feedbacks collected from users to learn the relationships among images. If there are too many combinations of interpretations or groupings for a finite number of images, it will not be possible to collect a sufficient number of user feedbacks, since each feedback would correspond to only one of such groupings.

We hypothesize that all the above assumptions are valid. Since it is difficult to derive their proofs theoretically, we conduct a series of experiments that suggest their validity.

4 Proof of concept

In order to test our idea, we build a simple content-free image-retrieval (CFIR) system based on collaborative filtering. We collected a data of user judgments on an image set. We also defined a performance measure of image retrieval. Using the data and measure, we compared the system's performance with respect to the varying number of user feedback, as well as with that of traditional content-based image retrieval (CBIR) system.

4.1 Rényi's entropy-based collaborative filtering algorithm

In this experiment we used a collaborative filtering algorithm developed by Zitnick [21], which was derived by maximizing Rényi's entropy. Other representative algorithms are Bayes Net [6].

Suppose there are n images in the database, $\boldsymbol{X} = \{I_1, \ldots, I_n\}$. The variable x_i is a logical variable associated with I_i. We denote $x_i = 1$ when *i*-th image I_i is selected and $x_i = 0$ when I_i is not selected. The image retrieval problem is to predict the probability of $x_i = 1$ given an observed condition, such as $\boldsymbol{X}_E = \{x_1 = 1, x_2 = 0\}$, which means I_1 is selected and I_2 is not selected by a user so far. We call such a condition set $\boldsymbol{X}_E$ an *evidence set*. Thus an image retrieval problem is computing $P(x_i = 1|\boldsymbol{X}_E)$ for all x_i that are not included in $_E$. In subsequent discussion, a notation for $\boldsymbol{X}_E$ is omitted, when it is obvious, to avoid clutter.

Since the possible combinations for $\boldsymbol{X}_E$ are huge, there will not be enough data to estimate for all $P(x_i = 1|\boldsymbol{X}_E)$. Yet, Zitnick showed that maximizing Rényi's entropy results in a good estimation of $P(x_i = 1|\boldsymbol{X}_E)$ as a weighted sum of functions $F = \{f_0, \ldots, f_c\}$. Each of f_i is a certain logical functions of $\{x_1, \ldots, x_n\}$.

(1) $$P(x_i = 1|\boldsymbol{X}_E) \sim \sum_j \lambda_{ij} f_j(\boldsymbol{X}_E)$$

λ_{ij} are Lagrange coefficients and they satisfy the following conditions.

(2) $$\lambda_{i\cdot}^T = \boldsymbol{p}_i^T \mathbf{P}^{-1}$$

$$\boldsymbol{p}_i = \begin{bmatrix} P(x_i = 1|f_0(\boldsymbol{X}_E)) \\ \vdots \\ P(x_i = 1|f_c(\boldsymbol{X}_E)) \end{bmatrix} \tag{3}$$

$$\mathbf{P} = \begin{bmatrix} P(f_0|f_0) & P(f_0|f_1) & \cdots & P(f_0|f_c) \\ \vdots & \vdots & \ddots & vdots \\ P(f_0|f_0) & P(f_0|f_1) & \cdots & P(f_0|f_c) \end{bmatrix} \tag{4}$$

$P(f_i|f_j)$ denotes $P(f_i(\boldsymbol{X}_E) = 1|f_j(\boldsymbol{X}_E) = 1)$ for all $\boldsymbol{X}_E$.

We set $f_0(\boldsymbol{X}_E) = 0$ and $f_i(\boldsymbol{X}_E) = (x_i = 1|\boldsymbol{X}_E)(i = 1, \ldots, n)$. The pair-wise conditional occurrence probability matrix **P** is estimated from the data.

A solution for $\boldsymbol{\Lambda} = [\lambda_{ij}]$ is,

$$\boldsymbol{\Lambda} = \mathbf{P}_{\cdot,E}\mathbf{P}_E \tag{5}$$

$\mathbf{P}_{\cdot,E}$ denotes a matrix whose *i*-th column is equal to *i*-th column of **P** if $xi \in \boldsymbol{X}_E$ or all zero column vector if $x_i \in \boldsymbol{X}_E$. $\mathbf{P}_E$ is a matrix whose element at *i*-th row and *j*-th column is equal to the element of **P** at *i*-th row and *j*-th column if $x_i, x_j \in \boldsymbol{X}_E$, or zero otherwise. $\mathbf{P}_E^+$ denotes a generalized inverse matrix of $\mathbf{P}_E$.

4.2 Data collection of user feedback

To evaluate an image retrieval system, we need ground-truth user data, i.e., a collection of judgments by people on whether certain images are relevant to each other within a set of images. Ideally, the data should be obtained from actual usage history of a relevance-feedback system. Here, however, we prepared a special data collection program. Figure 1 shows the interface used for data collection of user feedback.

A set of 10,000 images were prepared drawn from the *Corel image library* as the underlying image database. The set consisted of 50 images from each of 200 vendor-defined categories, so that the contents are broad and their distribution is balanced. Fifteen (15) human subjects (mostly students) were recruited to perform the data collection sessions. In each task session, a small subset (roughly 50) of images were randomly chosen from the underlying database, and were presented to a subject sitting at a computer monitor screen, with one image highlighted as a *target image* (see Figure 1). The location at which the target image is shown is randomized. The subject was asked to group images that are "similar" to the target image and to each other. The similarity criterion or the number of similar images to be selected was *not* specified.

For each performed task, a record R is created, consisting of the displayed image set D, displayed order O, the target image I, and the user-selected image set S.

Five subjects performed total of 4010 task sessions. Each of the other 10 subjects performed 100 tasks, 1000 tasks in total. Note that since the selection of ~50 images to be used as D from 10,000 image set, the order O they are displayed, and the selection of target image I are all randomized, there is no exact same task among these 5010 task sessions. Also note that 5010 is an infinitesimally small fraction of ${}_{10000}C_{50}$, all the possible selections of D.

4.3 Evaluation procedure and performance measure

We defined a procedure and criterion to evaluate an image retrieval method using the user data collected in the previous section. For each entry of task data $R = \{D, O, I, S\}$, k images from the selected image set S are given to the system as a query set Q. If there are not enough images in S, $|S| \leq k$, then the session data is not used.

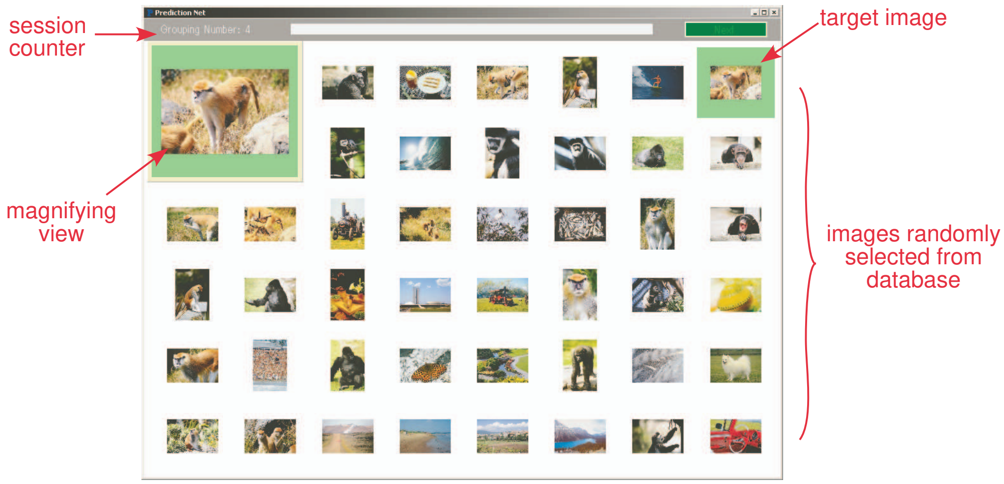

Figure 1: The interface for data collection of user feedback.

The image retrieval system ranks the images in D excluding the query images (i.e., images in $D - Q$). The accuracy of the ranking for the task R is defined as [21].

$$accuracy(\boldsymbol{R}) = \frac{\sum_{i=1}^{|D|-k} \delta(i, S) h(i)}{\sum_{i=1}^{|S|-k} h(i)} \tag{6}$$

$$h(i) = 2^{\frac{i-1}{b-1}} \tag{7}$$

where $\delta(i, S) = 1$ if i-th ranked image is in S, otherwise 0. $|D|$ and $|S|$ denote the number of images in D and S respectively. The value b is called "half-life" for $h(i)$, that is, $h(b) = 0.5$. Here, we used $b = 2$. The $accuracy(\boldsymbol{R})$ will be 1 when all images in $S - Q$ are ranked on top.

The assumptions behind this measure are the following. When using an image retrieval system, if a user submits one of images in S as a query and receives a subset of D including some images from S, the user will most likely select the images from S as relevant. Also, if the user receives only images from S in response to the query, the user will be most satisfied.

Finally, all $accuracy(\boldsymbol{R})$ are averaged over the entire test data to compute accuracy for the data set.

$$accuracy = \sum_i accuracy(\boldsymbol{R}_i) \tag{8}$$

4.4 Results

The described method was applied to evaluate our collaborative-filtering based content-free retrieval system as well as a typical content-based system that uses color coherent vector which is a combination of two color histograms [4].

For the collaborative-filtering based system, the 4010 task records from the first five users were used as training data, and the latter 1000 records were used as test data. The content-based system was tested with the same 1000 records. For k, we set the values at 1,2,5, and 10. Note that not all of the above 1000 test data contained enough number of selected images for the evaluation. Table 4.4 lists the number of usable test data records for each value of k.

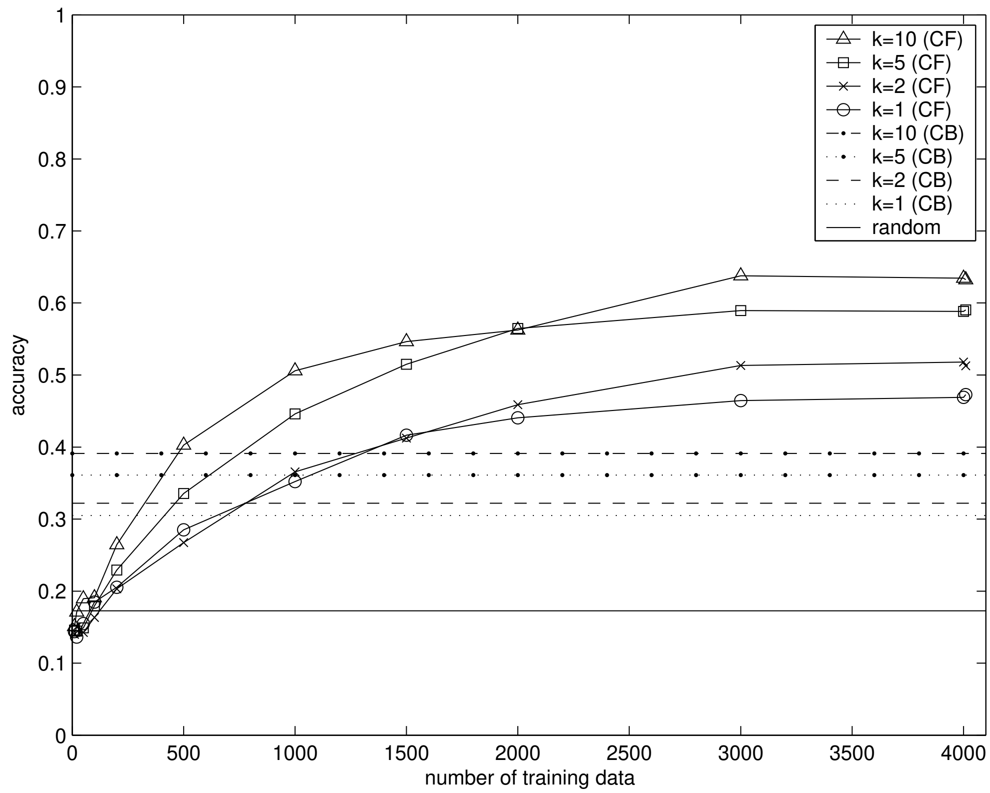

Figure 2: Image retrieval performance with respect to the number of training data and the number of sample images.

k	1	2	5	10
# of valid data	1000	784	462	265

Table 1: Number of usable test data records for each k.

The results were summarized in Figure 2. It plots the performance measure values with respect to the number of training data (100, 500, 1000, 1500, 2000, 3000, and 4010) for different numbers of sample images given as a query (k=1, 2, 5, and 10). The thin horizontal solid line at 0.173 corresponds to the accuracy the system would achieve if it returns random ranking. The four broken lines between 0.3 and 0.4 are the performance of the content-based image retrieval given different number of examples k=1, 2, 5, and 10, respectively.

The results clearly show that the performance of the content-free retrieval system improves as the number of feedback data increases. This indicates that the judgments on image relations made by one group of users helps another group of users, and suggests that their decisions more or less agree with each other. In other words, there is good inter-subject transferability of interpretation.

It is noteworthy that our collaborative filtering method trained with more than 1,000 feedback data already outperforms the content-based retrieval method for the task of image retrieval from 10,000 images. The result indicates the collaborative filter trained with 1500 feedback data and a single sample image

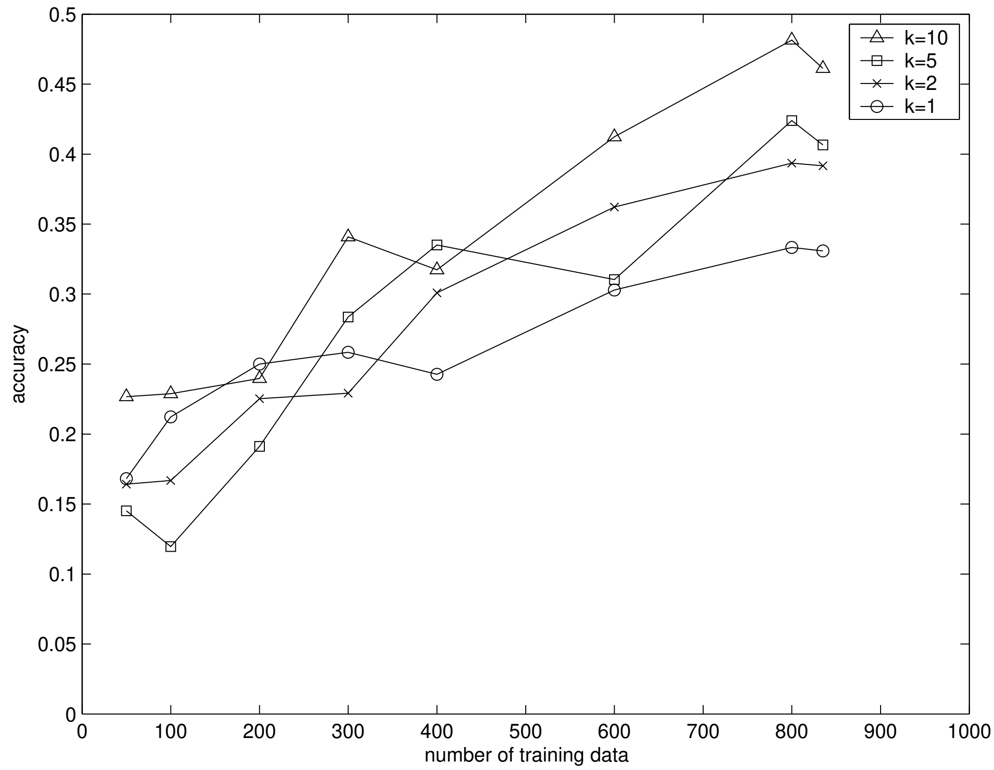

Figure 3: Retrieval performance using self-produced feedbacks.

performs as well as the tested content-based retrieval method with 10 example images[1]. Recall that the training data and the test data are from different users, and that the test data do not include any same task as in the training data. So, the content-free retrieval system worked well, not because it was given the same problems as those in training.

It is expected that intra-subject transferability is higher than inter-subject transferability; that is, if the training set and test set are drawn from the feedback data of the same subject, the training will be faster and better. Figure 3 shows the result of the intra-subject cases. Like in Figure 2, different numbers of session records of a *single* subject were used as training data, and 100 records of the same subject (but not the records included in the training set) were used as test data. Figure 3 shows that the accuracy values of an intra-subject case are comparable or higher than inter-subject cases. Consistency of interpretation is much higher within the same person as expected.

Some concerns remain with the result. $Accuracy$ curves appear to flat out as the number of training data increases. We do not know the exact reason for this phenomenon yet.

[1]We provided all the sample images to the systems at the same time. The result may have been different if the images were given incrementally to the content-based method.

5 Discussions

The experimental results appear promising, but still are very preliminary. In this section we discuss a few critical issues that need to be investigated further.

5.1 Cold start problem

Collaborative filtering is a "cold start" solution. The system needs be used for a certain amount of time before it accumulates enough data for learning and becomes capable. The more it is used, the more capable it becomes, but users may not want to use it unless it is capable. Another related problem is how to handle new images that have been added to the database.

One way to alleviate these difficulties is to use current text or content-based image retrieval techniques in combination. Similarities between images computed by content-based methods can be used to initialize the collaborative filtering. Also, a somewhat sneaky method to deal with new images is to insert them randomly in retrieval results as part of operation whenever the system retrieves images, and see how a user reacts to it.

Indeed, though we have emphasized the "content-free" aspect, we envision the final system to be a hybrid system. The collaborative filtering network is supported by other techniques that utilize any information associated with images, including content-based module and text-based module. A content-based module with relevance feedbacks is used initially and switches to a content-free method when enough user feedbacks are collected. For a web-wide image retrieval problem, statistical and thesaurus analysis of the textual information, such as caption or file name, which is often associated with images should be combined with collaborative filtering.

5.2 Number of feedbacks

It is interesting to know how many feedbacks are required to make our method work as intended.

In our algorithm, we need to estimate a pair-wise conditional occurrence probability matrix **P** in Equation (4). That is, probabilities for $\frac{N(N-1)}{2} \approx \frac{N^2}{2}$ pairs of images have to be computed where N is the total number of images in the database. It should be noted that most of the probabilities are zeros since each image is likely to be related to very limited portion of the database. Suppose each image is related to αN images, where $0 \leq \alpha \leq 1$, the number of probabilities that has to be estimated is $\frac{\alpha^2 N^2}{2}$.

Let N_D denote the number of images a user sees on the screen during one session when providing a feedback. Let also N_S denote the number of images that the user selects as relevant from the N_D images shown. We consider the rest of $N_{NS} = N_D - N_S$ images are implicitly labeled as non-relevant. From this one feedback session, $\frac{N_S(N_S-1)}{2} + N_S N_{NS} \approx N_S N_D - \frac{N_S^2}{2}$ pair-wise samples of image relations are obtained. Since interpretation and preference may differ between people, let us assume that among N_S image labeled as relevant, only βN_S images $(0 \leq b \leq 1)$ will have agreement with other people's interpretation. The final number of valid or usable relations are $\beta(N_S N_D - \beta\frac{N_S^2}{2})$. We assume β to be fairly large on the basis of our experiments that people's perception is more or less comparable. Using these notations, we can now discuss the number of feedbacks required to make our system work.

Suppose we need at least s feedbacks for reliable estimation of each pair-wise conditional probability. The total number of feedbacks F that are required is:

$$F_m < F < F_M \tag{9}$$

(10) $$F_m = \frac{s}{\beta N_S N_D - \frac{N_S^2}{2}} \left(\frac{\alpha N}{\beta}\right)^2$$

(11) $$F_M = \frac{s}{\beta N_S N_D - \frac{N_S^2}{2}} \left(\frac{N}{\beta}\right)^2$$

F_M corresponds to the case when images shown for the feedback session are completely randomly chosen, and F_m to the case when they are chosen most wisely (i.e., only those that need to be related are chosen).

If we set $N = 10000$, $N_S \approx 5$, $N_D \approx 50$, $\alpha = 0.01$, $\beta = 0.8$ and $s = 100$ (roughly corresponding to our experiment), we have $F - M = 8.3 \times 10^7$ and $F_m = 8.3 \times 10^3$.

A popular Internet search engine Google claims that it has indexed more than 425,000,000 images [23]. For this case, we have $F_m = 3.8 \times 10^{10}$, assuming each image is related with 2×10^5 images ($\alpha = 0.05\%$). Considering that the Google also answers 10^8 search queries per day, 3.8×10^{10} feedbacks can be collected roughly within a year.

The question of how to find the right pairs still remains. However, the problem is not unique to our method; it is the task for all image retrieval methods to find a small number of related images from a vast number of images. One of the advantages of our scheme is that once a set of images is identified as related by any mean, the knowledge is stored and reused. Because of this property, we expect the performance of our method to increase fairly monotonically as feedbacks accumulate.

5.3 Standard Data

While designing procedures and measures for evaluating image retrieval systems, we came to realize the strong need for their standardization, especially, the need for a standard corpus of images as we have seen in other research areas such as speech recognition and face recognition. Once we have a standard image set and associated user data set, researchers can refer to and use the same data to compare the performance directly.

We are planning to continue collecting user data and make them publicly available for the research community. Our current image data set is a commercial library that may not be distributed freely along with the user data. Although non-trivial, we are contemplating to build a free (or minimum-cost) open-source large-scale non-biased image dataset, following activities in other areas such as Open Video Library [5].

6 Summary and Conclusions

This paper has argued that having users in the loop is the key to capable image retrieval since current automated image understanding techniques have very limited capabilities.

We proposed a new user-powered "content-free" approach to image retrieval that directly utilizes and recycles feedbacks from users by means of collaborative filtering without doing image analysis. Recycling feedbacks not only reduces the burden of the users in providing the same information repeatedly, but also allows for accumulating the results of human's perceptual decisions on images.

The results of our preliminary experiment shows that the performance of "content-free" image retrieval system improves with the number of accumulated feedbacks, even outperforming a basic but typical traditional "content-based" system. Although many issues remain to be explored, our ultimate goal is to collect all computational powers from resources spread over networks both in time and space to accomplish a large-scale image retrieval task. The resources are human users.

Acknowledgments

We would like to thank Larry Zitnick for inspiring discussions and his collaborative filtering program. Many thanks go to Tat-Seng Chua and Zhao Yunlong of the National University of Singapore for providing their image retrieval program.

References

[1] Ahn, L. and Dabbish, L., "Labeling Images with a Computer Game," in *Proceedings of ACM CHI 2004*, pp. 319-326, 2004.

[2] Barnard, K. and Forsyth, D., "Learning the Semantics of Words and Pictures," in *Proceedings of International Conference on Computer Vision*, Vol 2, pp. 408-415, 2001.

[3] Chang, N.-S. and Fu, K.-S., "Query by pictorial example," in *IEEE Transactions on Software Engineering*, Vol. 6, No. 6, pp. 519-524, 1980.

[4] Chua, T.-S. and Chu, C.-X., "Color-based pseudo object model for image retrieval with relevance feedback," in *Proc. of First International Conf. on Advanced Multimedia Content Processing*, pp. 145-160, 1998.

[5] Geisler, G. and Marchionini, G., "The Open Video Project: A Research-Oriented Digital Video Repository," in *Proc. of ACM Digital Libraries*, pp. 258-259, 2000.

[6] Heckerman, H., Geiger, D., Chickering, D. M., "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, Vol. 20, No. 3, pp. 197-243, Kluwer Academic Publishers, 1995.

[7] Horowitz, B., Humphrey, R., Jain, R., Shu. CF., "Virage image search engine: An open framework for image management," in *Proc. of SPIE Conf. on Storage and Retrieval for Image and Video Databases*, pp. 76-87, 1996.

[8] Kumar, S., Hebert, M., "Man-made structure detection in natural images using a causal multiscale random field," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 119-126, 2003.

[9] Lipson, P., Grimson, E., Sinha, P., "Configuration Based Scene Classification and Image Indexing," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1007-1013, 1997.

[10] McNee, S. M., Lam, S. K., Konstan, J. A., Riedl., J., "Interfaces for Eliciting New User Preferences in Recommender Systems," in *Proc. of The 9th International Conf. on User Modeling (UM'2003)*, pp. 178-188, 2003.

[11] Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Pektovic, D., Yanker, P., Faloutsos, C., Taubin, G., "The QBIC project: querying images by content using color, texture, and shape," in *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases V*, Vol. 1908, pp. 173-187, 1993.

[12] Rowley, H. A., Baluja, S., Kanade, T., "Neural Network-Based Face Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 23-38, 1998.

[13] Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., "Relevance feedback: a power tool for interactive content-based image retrieval," in *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 8, No. 5, pp. 644-655, 1998.

[14] Smeulders, A. W. M., Woming, S., Santini, S., Gupta, A., Jain, R., "Content-Based Image Retrieval at the End of the Early Years," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1349-1380, 2000.

[15] Smith, J. R. and S. F. Chang, S.-F., "An Image and Video Search Engine for the World Wide Web," in *Proc. of SPIE Conference on Storage and Retrieval for Image and Video Databases*, Vol. 3022, pp. 84-95, 1997.

[16] Szummer, M. and Picard, R. W., "Indoor-Outdoor Image Classification," in *Proceedings of IEEE Workshop on Content-based Access of Image and Video Databases*, pp. 42-51, 1998.

[17] Vailaya, A., Jain, A., Zhang, H. J., "On image classification: city images vs. landscapes," in *Pattern Recognition*, Vol. 31, No. 12, pp. 1921-1935, 1998.

[18] Wang, J. Z., Li, J., Wiederhold, G., "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 9, pp. 947-963, 2001.

[19] Wang, J. Z. and Li, J., "Learning-Based Linguistic Indexing of Pictures with 2-D MHMMs," in *Proc. of ACM Multimedia*, pp. 436-445, 2002.

[20] Zhou, X. S., Rui, Y., Huang, T. S., *Exploration of Visual Data*, Kluwer Academic Publishers, 2003.

[21] Zitnick, C., "Computing Conditional Probabilities in Large Domains by Maximizing Rényi's Quadratic Entropy," *doctoral dissertation*, tech. report CMU-RI-TR-03-20, Robotics Institute, Carnegie Mellon University, May, 2003.

[22] http://www.amazon.com

[23] http://www.google.com