CONTENT-FREE IMAGE RETRIEVAL BASED ON RELATIONS EXPLOITED FROM USER FEEDBACKS

*Shingo Uchihashi** and *Takeo Kanade*

Department of Electrical and Computer Engineering* and Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, USA
{shingo,tk}@cs.cmu.edu

ABSTRACT

We propose a new "content-free" image retrieval method which attempts to exploit certain common tendencies that exist among people's interpretation of images from user feedbacks. The system simply accumulates records of user feedback and recycles them in the form of collaborative filtering. We discuss various issues of image retrieval, argue for the idea of content-free, and present results of experiment. The results indicate that the performance of content-free image retrieval improves with the number of accumulated feedbacks, outperforming a basic but typical conventional content-based image retrieval 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. Many efforts have been made in the last decade [1], and they reveal that a key to a capable image retrieval system is how to extract and describe the image contents.

One obvious approach is to describe the image contents verbally, typically keywords. Once the verbal descriptions are obtained, text search techniques can be applied to retrieve images in the database allowing query-by-keyword. However, this assumption is seldom met; manual labeling is too expensive and automatic methods are not reliable for the moment. Limited success is reported in automatic image classification [2]. Only few objects, such as faces or cars can be recognized reliably from general images. Recent attempts to automatically learn the relations between image regions and keywords have not yet achieved satisfactory results [3]. Some researchers turned to alternative information sources. For images on web pages, the use of file names, path names, and surrounding text has been proposed and deployed by commercial search engine companies. Ahn *et al* [4] has proposed a novel approach to combine manual image labeling and network games.

The other approach is to represent images with nonverbal descriptions which can be reliably computed from images. Typical such descriptions are image features based on color, shape, and texture[1]. Conventional content-based image retrieval (CBIR) methods use these image features to define image similarity. Finding a good set of features is very critical since the rest is built upon it. However, since we have not revealed human visual perception mechanisms, proposed features and image similarity measures are rather computer-centric. Interestingly, even such features work reasonably well in some occasions, although they achieve severely limited success for most of cases. 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. Human perception of images is complex and seems to be dependent on context, purpose, and individual cases. Image representations need to reflect such characteristics of human visual perception.

Besides seeking for more suitable image features, many researchers have reported that improved results are obtained by incorporating user feedbacks into the content-based image retrieval system [5][6]. 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 the query. Given relevance 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.

We have proposed a new approach to image retrieval that uses user feedbacks in the form of interpretation rather than through image features, thus directly utilizing human perceptive power [7]. Relations among images are exploited rather than the image "contents". 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 performance. We will name our approach "content-

0-7803-9332-5/05/$20.00 ©2005 IEEE

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. CONTENT-FREE IMAGE RETRIEVAL

2.1. Content-free Concept

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 human can provide consistent and reliable judgment of whether presented images are relevant to what he/she is looking for. By receiving the teaching signals, content-based methods can learn how to respond to the query. However, we observe two different types of limitation in this scheme. Firstly, the selection of image feature limits the capability of model-fitting. Secondly, several iterations of feedbacks will not provide enough data to train a complex vision model. To utilize knowledge from users more effectively, we omit image features and use the human's perceptual decisions themselves.

Note that relevance feedbacks are tolerable amount of manual labor enforced on users to achieve their goal. Because of this nature, each feedback carries little but reliable information regarding how images are related to each other. We believe that an effective image retrieval system can be realized using only the usage history of users. We record all of these feedbacks from all of the users. The aggregated feedbacks should work as asynchronous voting on relations among images in the database. Once enough feedbacks are accumulated, the system can learn and summarize those relations in a certain form. Subsequently the system retrieves relevant images for a new query from a new user using the learned relations, 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 relations from them to do the retrieval task. Hence the name: "content-free" approach.

Some research efforts have been conducted in similar concept. However, they used the accumulated user feedbacks in content-based frameworks so that the performance is restricted by image features [8][9].

2.2. 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. We use a collaborative filtering algorithm developed by Zitnick [10]. Other representative algorithms such as Bayes Net are also applicable.

Suppose there are n images in the database. The variable $x_i \in \boldsymbol{X}$ is a logical variable associated with image 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 $\boldsymbol{X}_E = \{x_1 = 1, x_2 = 0\}$. We call such a condition set $\boldsymbol{X}_E$ an *evidence set*. More formally, image retrieval problem is computing $P(x_i = 1|\boldsymbol{X}_E)$ for all $x_i \in \boldsymbol{X}_H$ where $\boldsymbol{X}_H = \boldsymbol{X} - \boldsymbol{X}_E$. 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 $P(x_i = 1|\boldsymbol{X}_E)$. Zitnick showed that by maximizing Rényi's entropy, the best estimation of $P(x_i = 1|\boldsymbol{X}_E)$ using $F = \{f_0, \ldots, f_c\}$, which is a set of functions of $\{x_1, \ldots, x_n\}$, is obtained as a weighted sum of the functions, that is,

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

where λ_{ij} are Lagrange coefficients whose values can be computed from $\boldsymbol{X}_E$ and the pair-wise conditional occurrence probability matrix **P** below. Note **P** can be estimated from the accumulated user feedbacks. Thus, the estimate of $P(x_i = 1|\boldsymbol{X}_E)$ can be always computed. See [10] for more details.

$$\mathbf{P} = \begin{bmatrix} P(f_0|f_0) & P(f_0|f_1) & \cdots & P(f_0|f_c) \\ \vdots & \vdots & \ddots & \vdots \\ P(f_0|f_0) & P(f_0|f_1) & \cdots & P(f_0|f_c) \end{bmatrix} \qquad (2)$$

$P(f_i|f_j)$ denotes $P(f_i(\boldsymbol{X}_E) = 1|f_j(\boldsymbol{X}_E) = 1)$ for all $\boldsymbol{X}_E$. We set $f_0(\boldsymbol{X}_E) = 1$ and $f_i(\boldsymbol{X}_E) = (x_i = 1|\boldsymbol{X}_E)$.

3. PROOF OF CONCEPT

In order to test our idea, we build a simple CFIR system based on the above collaborative filtering algorithm. We conducted a series of experiments to verify the basic concept.

3.1. Data Collection of User Feedback

A collection of judgments by people on whether certain images are relevant to each other within a set of images is required to train a collaborative filtering system. Ideally, the data should be obtained from actual usage history of a relevance-feedback system. Here, however, we prepared a special data collection program to facilitate the process.

A set of 10,000 images were prepared from the *Corel image library* consisting of 50 images from each of 200

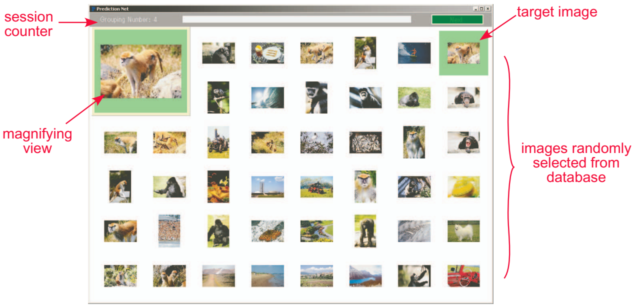

Fig. 1. The interface for data collection of user feedback.

vendor-defined categories. Fifteen subjects were asked to form a group of images from 44 images on the screen, one of which is highlighted to indicate the target image (See Figure 1). The only instruction given was to select images "similar" to the target image and to each other. The similarity criterion or the number of similar images to be selected was *not* specified. Our data collection program is designed to imitate content-based image retrieval processes. The displayed images are intended to be initial retrieval results where a user is seeking images represented by the associated target image.

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.

3.2. Evaluation Procedure and Performance Measure

We evaluate our 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 (or $\mathbf{X}_E$). If there are not enough images in S, $|S| \leq k$, then the session data is not used.

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 [10].

$$accuracy(\boldsymbol{R}) = \frac{\sum_{i=1}^{|D|-k} \delta(i, S) h(i)}{\sum_{i=1}^{|S|-k} h(i)} \qquad (3)$$

where $h(i) = 2^{i-1}$ and $\delta(i, S) = 1$ if i-th ranked image is in S, otherwise 0.

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.

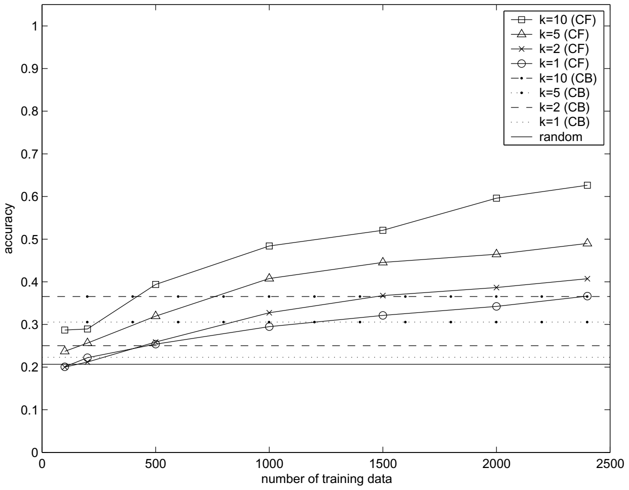

Fig. 2. Image retrieval performance with respect to the number of training data and the number of sample images.

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

$$accuracy = \sum_{i} accuracy(\boldsymbol{R}_i) \qquad (4)$$

3.3. Results

We evaluated the performance of our system using 2500 user feedback data from 25 subjects in leave-one-out scheme, that is, we tested how well every 100 records from a single subject can be predicted using records from the rest of 24 subjects as training data. The described method was applied to evaluate our collaborative-filtering based CFIR system as well as a typical color-based CBIR system described in [5].

Figure 2 summarizes results of our expriment. Different numbers of sample images were given as queries (k=1, 2, 5, and 10). The performance of CFIR system is also compared with a CBIR method and random ranking in Figure 2.

The results clearly show that the performance of the content-free retrieval system improves as the number of feedback data provided 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.

4. 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.

4.1. Cold Start Problem

Collaborative filtering is a "cold start" solution. The system needs some leadoff time to accumulates enough data before starting to produce meaningful output, but users may not want to use it unless it provides anything useful. Related to this topic, how to handle new images that have been added to the database is another issue.

One way to alleviate these difficulties is to use current text or CBIR techniques in combination. Similarities between images computed by content-based methods can be used to initialize the collaborative filtering. 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.

4.2. 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 (2). That is, $\frac{N(N-1)}{2} \approx \frac{N^2}{2}$ numbers have to be computed where N is the total number of images in the database. Considering the diversity of the large-scale image database, most of the probabilities are zeros. Therefore, the number of probabilities that have to be estimated is αN^2, where $0 \leq \alpha \ll 1$.

A popular internet search engine Google claims that it has indexed more than 425,000,000 images. Let us assume each image has *one million* related images, i.e., $\alpha = 6.25 \times 10^{-6}$. Google assumably answers 3×10^7 image search queries per day. For a typical session, 20 images are presented on the screen at once. If each user clicks five images per session, providing $\sim 5 \times 20 = 100$ image relations, sufficient number of feedbacks to estimate the probability matrix can be collected roughly in one year.

Although the question of how to find the right pairs still remains, our scheme has an advantage because once a set of images is identified as related by any mean, the knowledge is stored and reused.

5. CONCLUSIONS

In this paper, the cocept of our "content-free" approach to image retrieval is demonstrated through experiments. By accumulating user feedback data, relations among images are obtained as a conditional probability matrix that is used in a collaborative filtering algorithm. The results show that our scheme can achieve high retrieval performance without analyzing image contents.

Content-free approach is a new frontier to image retrieval. Many issues remain to be explored as well as many possibilities. We focused on user feedbacks in this paper, but there are other bits of reliable information available including text in web pages that are already exploited in other systems. 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.

6. REFERENCES

[1] A. W. M. Smeulders, S. Woming, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: city images vs. landscapes," *Pattern Recognition*, vol. 31, no. 12, pp. 1921–1935, 1998.

[3] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proceedings of Int. Conf. on Computer Vision*, 2001, pp. 408–415.

[4] Luis von Ahn and Laura Dabbish, "Labeling images with a computer game," in *Proceedings of ACM CHI 2004*, 2004, pp. 319–326.

[5] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. AMC Int. Multimedia Conf.*, 2001, pp. 107–118.

[6] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, 1998.

[7] Takeo Kanade and Shingo Uchihashi, "User-powered "content-free" approach to image retrieval," in *Proceedings of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society*, 2004, pp. 24–32.

[8] H. Möler, T. Pun, and D. Squire, "Learning from user behavior in image retrieval: Application of the market basket analysis," *International Journal of Computer Vision*, vol. 56, no. 1–2, pp. 65–77, 2004.

[9] X. He, O. King, W.-Y. Ma, M. Li, and H.-J. Zhang, "Learning a semantic space from user's relevance feedback for image retrieval," *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 39–48, 2003.

[10] Charles Zitnick, *Computing Conditional Probabilities in Large Domains by Maximizing Rényi's Quadratic Entropy*, Ph.D. thesis, Robotics Institute, Carnegie Mellon University, May 2003, Technical Report CMU-RI-TR-03-20.