Content-Free Image Retrieval by Combinations of Keywords and User Feedbacks

Shingo Uchihashi[1] and Takeo Kanade[2]

[1] Department of Electrical and Computer Engineering
[2] Robotics Institute,
Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh PA 15213, USA
{shingo, tk}@cs.cmu.edu

Abstract. The performance of a new content-free approach to image retrieval is demonstrated. Accumulated user feedback data that specify which images are (ir)relevant to each other and keywords obtained from a network game are recycled through collaborative filtering techniques to retrieve images without analyzing actual image pixels. Experimental results show the proposed method outperforms a conventional content-based approach using support vector machine. The result was achieved by the combination of feedback data and keywords. Applications of the proposed scheme in query-by-text image retrieval is also discussed.

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. Image retrieval systems have to know all the interpretations in order to respond queries from users. However, image interpretation is a highly complicated perceptive process and currently only human can perform the task.

Conventional content-based image retrieval (CBIR) methods deploy computer-centric representations of images for automated image indexing. Typically, statistical characteristics of pixel values or patterns in color, shape, and texture composition in an image are used as image features. *Similarity* between images is computed based on the image features. While this relatively simple scheme has achieved certain success, its performance is severely limited because of the "semantic gap", the difference between what image features represent and what people perceive from the image.

Authors have introduced a new approach to image retrieval that directly utilizes human's perceptual capability [6]. From evidences of how people perceive image contents, our method reproduces human's judgments on the contents. In our previous work, we used relevance feedbacks as such evidences. It has been shown that relevance feedbacks from users can improve the performance of CBIR. From output images the system produces, a user specifies which images are (ir)relevant to what the user desires to retrieve, and the system adjust the internal parameters of similarity functions to adapt the individual user according to the feedbacks. This is indeed one way to incorporating human's perceptial capability to CBIR. However, the use of feedbacks in the CBIR framework

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 650–659, 2005.

is limited by the "semantic gap". Instead of processing images, we simply collect user feedbacks to directly exploit human perceptive power and certain common, if not identical, tendencies that must exist among people's interpretation and preference of images. To illustrate the point, we call our approach "content-free" image retrieval (CFIR).

In this paper, we explore the use of keywords as evidences of human's image perception in addition to relevance feedbacks. Keywords are more explicit form for users to express image contents than relevance feedbacks. Because it also takes more labor, it has been considered to be expensive and unrealistic to ask people to manually provide keywords to each image in a large-scale image database. Recently Ahn *et al* [14] has proposed an interesting approach to make manual image labeling into a network game, in which game participants are led to willingly do the labeling task. We obtained keywords from this novel scheme and used a collaborative filtering technique to integrate keywords and user feedback data for our CFIR system. The retrieval performance is computed and compared with a standard CBIR scheme.

2 Related Works

Image retrieval has been an active research area for the last decade [10, 18]. It started as a natural extension of document retrieval. Image contents were described using text, typically keywords, and simply text retrieval techniques were applied to retrieve text and associated images. The difficulty with this approach lies in how to get such text data. As manual labeling is too costly, alternative sources are necessary. Many attempts have been made to automatically classify images [13] or recognize objects in them [8]. Several methods have been proposed to learn the relationship between image regions of specific color or pattern, and keywords [1, 15]. 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.

Content-based image retrieval methods deploy computer-centric image descriptors, typically low-level image features, and therefore suffer from the semantic gap [16]. Various features and associated *similarity* measures have been proposed that attempt 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. No single set of features and similarity measure is applicable for all the cases. Adapting similarity measure for each query improves retrieval performance. Relevance feedback mechanisms with which users tell the system which images are (ir)relevant to what they want, are widely adapted into CBIR systems to adjust similarity measure computed from the image features. Many researchers have reported that improved results are obtained [9, 12].

While it looks promising we observe two different types of limitation in the current content-based methods with relevance 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. Recently, several research groups pointed out the insufficient data issue and proposed to accumulate feedbacks from the users in the past [4, 7]. Among them, the work by Möler *et al* is most similar to ours [7]. Like others, Möler uses the accumulated user feedbacks only to expand the current query and retrieval is done using a CBIR scheme, therefore the semantic gap still persists.

Zitnick and authors introduced content-free image retrieval [6, 19]. CFIR tries to take advantage of the fact that human users know the image contents. From products of human perception, CFIR seeks for the underlying knowledge. Similar ideas are explored in world-wide-web image retrieval research. From HTML descriptions, image file names, image path names, and alternative text for images are extracted and analyzed[2, 11]. Text surrounding images are used as additional information [17]. These kinds of information are manually attached to images either directly or indirectly, and therefore considered to be more accurately describing the image contents. The availability of information from web pages is limited and we cannot use it to actively index images. As mentioned earlier, Ahn *et al* proposed a novel image labeling scheme whose idea just matches with CFIR [14].

3 Integrating Keywords and User Feedbacks into Content-Free Formulation

3.1 Image Retrieval Problem

We formulate an image retrieval problem as follwoings. Here we consider query-by-example image retrieval to contrast with typical CBIR systems with relevance feedback. Suppose there are n images in the database, $\boldsymbol{I} = \{I_1, \ldots, I_n\}$. 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 $\boldsymbol{X}_E$. In subsequent discussion, a notation for $\boldsymbol{X}_E$ is omitted, when it is obvious, to avoid clutter.

3.2 Rényi's Entropy-Based Collaborative Filtering Algorithm

Since the possible combinations for $\boldsymbol{X}_E$ are huge, it is not realistic to estimate all $P(x_i = 1|\boldsymbol{X}_E)$ from data. Zitnick showed that by maximizing Rényi's entropy, a good estimation of $P(x_i = 1|\boldsymbol{X}_E)$ is obtained as a weighted sum of functions $F = \{f_0, \ldots, f_c\}$, where each of f_i is a certain logical functions of $\{x_1, \ldots, x_n\}$ [19].

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

λ_{ij} are Lagrange coefficients under the following constraints.

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

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

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

$P(f_i|f_j)$ denotes $P(f_i(\boldsymbol{X}_E) = 1|f_j(\boldsymbol{X}_E) = 1)$. We set $f_0(\boldsymbol{X}_E) \equiv 1$ and $f_i(\boldsymbol{X}_E) \equiv (x_i|\boldsymbol{X}_E)(i = 1, \ldots, n)$.

In (4), the pair-wise conditional occurrence probability matrix **P** can be estimated from user feedback data.

3.3 Combining Keywords and User Feedbacks

Zitnick's algorithm shown in the Section 3.2 is a general framework and its application is not limited to query-by-example image retrieval tasks. We apply the algorithm to a case where keywords are attached to images. Consider the word database $\boldsymbol{W} = \{W_1, \cdots, W_m\}$ which contains m keywords. K_i denotes a set of keywords attached to the image I_i. Each K_i is a subset of $\boldsymbol{W}$. By introducing logical variable y_j, which is associated with W_j, and allowing $\boldsymbol{X}_E$ to include y_j's, we can derive the same approximation as in the previous section. Here we have $f_{(n+j)}(\boldsymbol{X}_E) \equiv (y_j|\boldsymbol{X}_E)(j = 1, \ldots, m)$.

With our arrangement, the conditional probability matrix **P** is splitted into four sections as in (5). Note that we omitted the first row and the first column of **P** in (5) to simplify the argument, however they are still reserved as described in Section 3.2.

$$\mathbf{P} = \left[\begin{array}{c|c} \mathrm{P}_1 & \mathrm{P}_2 \\ \hline \mathrm{P}_3 & \mathrm{P}_4 \end{array}\right] \tag{5}$$

P_1 defines pair-wise image-to-image relations and is equivalent to (4). P_2 and P_3 represents how words are attached to images. We set $\mathrm{P}_2(f_i|f_{(n+j)}) = \frac{1}{tf_j}$, where tf_j is the term frequency of y_j, and $\mathrm{P}_3(f_{(n+j)}|f_i) = 1$ when $y_j \in K_i$. P_4 defines pair-wise word-to-word relations. Setting P_4 as an identity matrix results in the exact word matching. P_4 may be pre-computed using a dictionary to allow expanded word matching. One of the advantages of text data is that we can use additional resources such as grammers and dictionaries [18].

4 Comparison of CFIR and CBIR

In this section, we compare retrieval performance of our CFIR algorithm and a standard CBIR method. A set of 10,000 images were 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. The performance two systems are compared by *precision* using the vendor-defined categories as the ground truth.

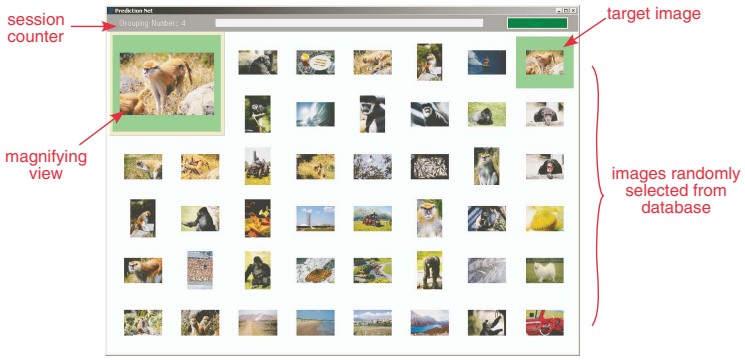


Fig. 1. The interface for data collection of user feedback

4.1 Collecting User Feedback Data and Keywords

User Feedback Data. We collect user feedback data as evidences of how people judge image contents to compute **P**. Ideally, the feedback data should be obtained from actual usage history of a relevance-feedback system. Here, however, we prepared a tailored data collection program. Having human subjects in controlled environments allowed us to collect data systematically and thus to facilitate the data collection process. Figure 1 shows the interface used for data collection of user feedback.

25 human subjects (mostly students) were recruited to perform the data collection sessions. 44 images were chosen uniformly randomly from all the 10,000 images, and displayed to a subject. A participating subject is asked to sit at a computer monitor screen. Among the displayed images, one image is 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. Each subject conducted 100 sessions.

One may interpret that our data collection software is simulating a dumb image retrieval system which simply returns randomly selected results.

Keyword Acquisition. We adopted the ESP Game to collect keywords for the images [14]. It is a network game which leads participants to willingly do the image labeling task. We registered our images with the ESP Game system and collected the total of 64,131 words in three months period. The maximum number of keywords per image is 41 and the minimum is 1. The keywords were lemmatized using WordNet and words not registered as nouns in the dictionary were omitted [3]. The vocabulary size of the ESP words is 3,073.

4.2 Reference CBIR (SVM) Implementation

We implemented a CBIR system using Support Vector Machine (SVM) as described in [12]. For image features, we used color correlograms following [5]. RGB color space is equally divided into $4\times4\times4$ bins. Four depth levels are used, $D = \{1, 3, 5, 7\}$. The

SVM uses a gaussian kernel with $\sigma = 0.1$ which we experimentally determined to produce the best performance for our images.

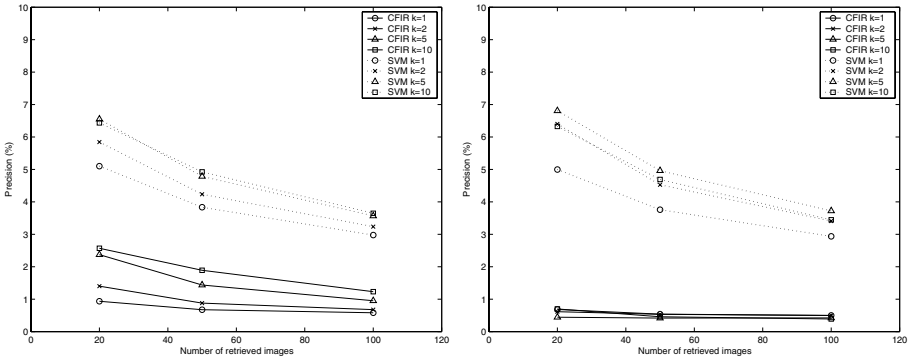
4.3 Sample Selection

For CFIR, sample images are selected following the procedures as below: 1) Randomly select k positive sample images from one category and k negative sample images from the rest. 2) Compute output for the selected samples. 3) Evaluate the performance by precision at scale=20, 50, 100. 4) Find k unlabeled positive samples from the top 100 ranked images, starting from the bottom, and label them. If not enough positive samples are available, label negative samples from the top of the ranking. Label k samples in total. 5) Iterate 2–4 two times. 6) Repeat the process for all the categories.

For CBIR, we followed the sample selection scheme described in [12]: 1) Randomly select k positive sample images from one category and k negative sample images from the rest. 2) Compute ouput for the selected samples. Rank images according to the distance from the decision boundary. 3) Evaluate the performance by precision at scale=20, 50, 100. 4) Label $k/2$ positive samples and $k/2$ negative samples from the 100 images closest to the boundary. 5) Iterate 2–4 two times. 6) Repeat the process for all the categories.

4.4 Preliminary Experiment

First, we compared the performance of the original implementation of CFIR described in 3.2 with the CBIR method. User feedback data was used to train the collaborative filter. Figure 2 (a) shows the precision curves of the initial outputs and Figure 2 (b) shows the results after the second iteration. Although the performance of our CFIR is slightly less than a half of the reference performance in Figure 2 (a), it is still encouraging because the performance is clearly better than a random selection ($\approx 0.5\%$) where we started from. We think the poor performance in Figure 2 (b) ($\approx 0.5\%$) is due to insufficient training data. If an image is not connected with other images then the image



(a) Initial output (b) After the second iteration

Fig. 2. Precision curve comparison between CFIR and CBIR(SVM)

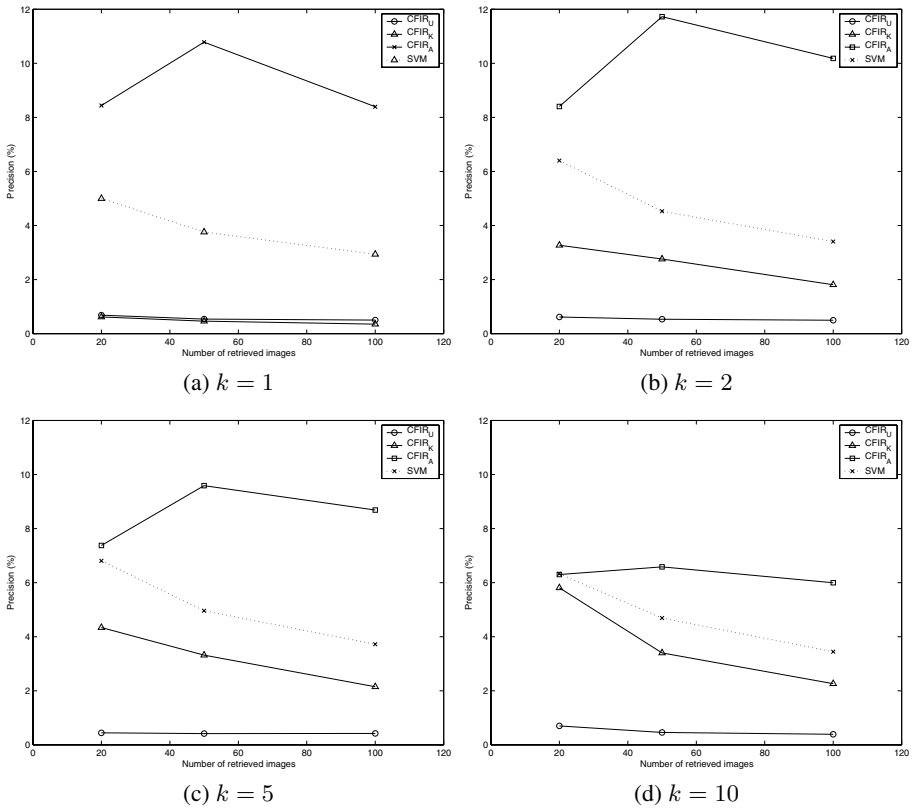
will not appear in the retrieval result because there is no supporting data. Our sample selection scheme tends to capture "connected" images and since the number of such images is small, there may be none left after two iteration.

4.5 Retrieval Performance Comparison

The following configurations were compared following the procedures described in the above.

- CFIR trained using user feedback data only (**CFIR**$_U$)
- CFIR trained using keywords only (**CFIR**$_K$)
- CFIR trained using user feedback data and keywords (**CFIR**$_A$)
- CBIR (**SVM**)

Note that CFIR$_U$ is the same as in Section 4.4. CFIR$_A$ uses the same **P** from Section 4.4 as P_1. P_1 in CFIR$_K$ is an identity matrix. For both CFIR$_K$ and CFIR$_A$, P_2 and P_3 of (5) were filled using the relations between keywords and images obtained through the ESP Game. P_4 was set to be the identity matrix.When we select images as



(a) $k = 1$ (b) $k = 2$

(c) $k = 5$ (d) $k = 10$

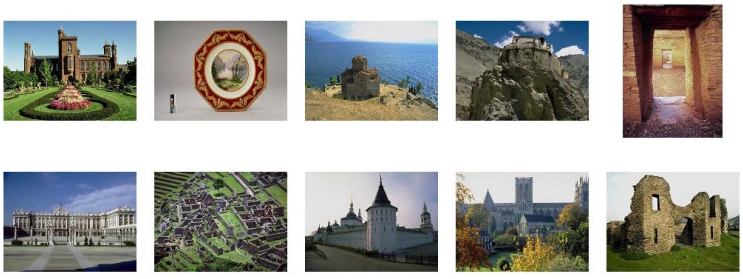
Fig. 3. Retrieval performance comparison

samples, partial scores are granted to the attached keywords depending on how common the word is among the selected samples. For example, suppose image I_1, I_2, I_3 are selected and the corresponding attached keyword sets are $K_1 = \{W_1, W_2\}, K_2 = \{W_2, W_3\}, K_3 = \{W_2, W_3, W_4\}$, then the evidence set for this sample set becomes $\boldsymbol{X}_E = \{x_1 = 1, x_2 = 1, x_3 = 1, y_1 = 0.3333, y_2 = 1.0, y_3 = 0.6667, y_4 = 0.3333\}$.

Figure 3 (a)-(d) show the precision curves of the four configurations after the second iteration for $k = 1, 2, 5, 10$. It is noteworthy that neither user feedback data nor keywords alone was enough to train CFIR to perform better than CBIR (See CFIR_U, CFIR_K, and SVM in Figure 3). But the combination of the two makes CFIR to achieve better result than SVM (See CFIR_A and SVM).

5 Alternative Query Modes in Content-Free Image Retrieval

So far, images and words are treated equally. We started from a query-by-example image retrieval system and then expanded it to incorporate with text information. Alternatively, a user can issue a query by first submitting keywords. That is, an evidence set $\boldsymbol{X}_E$ can containing only y_j can be composed to compute the corresponding output. For example, a sample query "Europe castle" is converted to $\boldsymbol{X}_E = \{y_{907} = 1(\text{Europe}), y_{445} = 1(\text{castle})\}$ and our algorithm described in Section 3 computes a



(a) Top 10 results for query "Europe castle"

(b) Top 10 results for query "brown dog"

Fig. 4. Sample outputs for query-by-text

probability for each image to be preferred given the query, $P(x_i = 1|\boldsymbol{X}_E)$. Figure 4 shows preliminary results of our system for query-by-text.

The system takes keywords as queries and shows the intial outputs to users. Then users provide feedbacks by clicking on the output images, and the system return the updated results. By accumulating the feedbacks and keywords in queries that are obtained through the interactions with users, the system can learn more about the images and produce better retrieval results in the future. Currently the vocabulary size is too small to produce practical outputs for a wide range of inputs. Integration with a full-scale dictionary will be necessary.

Further a different mode of operation is possible. One interesting topic is word estimation from images. By first providing images, corresponding keywords can be retrieved following an inversed path of query-by-text image retrieval, thus image contents may be estimated as words from sample images.

6 Summary and Conclusions

In this paper, we demonstrated the performance of a content-free image retrieval system. As evidences of how human perceive images, relevance feedback data were collected using a simulated environment. Also, descriptions of the image contents as keywords were collected through the ESP Game. We extended our previously proposed method to incorporate with keywords as well as user feedback data. The collected evidences were accumulated and recycled in the form of a collaborative filter.

Experimental results showed that the combined use of feedback data and keywords compensate achieve better retrieval performance than a standard content-based method using SVM. although each data alone was outperformed by the conventional scheme.

Applications of the proposed scheme in query-by-text image retrieval was also discussed in this paper with some preliminary results. Our algorithm treats images and keywords equally, therefore alternative usage modes are possible including image retrieval from keywords and keyword estimation from images.

Acknowledgments

We would like to thank Larry Zitnick for inspiring discussions and his collaborative filtering program, Luis von Ahn for kindly allowing us use the ESP Game to collect keywords for the images, and Satoshi Ichimura of Tokyo University of Technology for his support in recruiting the volunteers.

References

1. K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proceedings of International Conference on Computer Vision*, pages 408–415, 2001.
2. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

3. Christiane Fellbaum, editor. *WordNet: An Electric Lexical Database*. The MIT Press, May 1998.
4. 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 Transactions on Circuits and Systems for Video Technology*, 13(1):39–48, 2003.
5. Jing Huang, S Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.
6. 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*, pages 24–32, 2004.
7. 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*, 56(12):6577, 2004.
8. H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
9. Yong Rui and Thomas Huang. Optimizing learningg in image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 236–243, 2000.
10. 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 Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
11. J. R. Smith and S. F. Chang. An image and video search engine for the world wide web. In *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume 3022, pages 84–95, 1997.
12. S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proc. AMC Int. Multimedia Conf.*, pages 107–118, 2001.
13. A. Vailaya, A. Jain, and H. J. Zhang. On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935, 1998.
14. Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of ACM CHI 2004*, pages 319–326, 2004.
15. J. Z. Wang and J. Li. Learning-based linguistic indexing of pictures with 2-d mhmms. In *Proceedings of ACM Multimedia*, pages 436–445, 2002.
16. J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.
17. Xin-Jing Wang, Wei-Ying Ma, Gui-Rong Xue, and Xing Li. Multi-model similarity propagation and its application for web image retrieval. In *Proceedings of ACM Multimedia*, pages 944–951, New York City, USA, 2004.
18. X. S. Zhou, Y. Rui, and T. S. Huang. *Exploration of Visual Data*. Kluwer Academic Publishers, 2003.
19. Charles Zitnick. *Computing Conditional Probabilities in Large Domains by Maximizing R'enyi's Quadratic Entropy*. PhD thesis, Robotics Institute, Carnegie Mellon University, May 2003. Technical Report CMU-RI-TR-03-20.