A COLLABORATIVE BAYESIAN IMAGE RETRIEVAL FRAMEWORK

Rui Zhang, Ling Guan

Ryerson Multimedia Research Laboratory
Ryerson University, Toronto, Canada
{rzhang, lguan}@ee.ryerson.ca

ABSTRACT

In this paper, an image retrieval framework combining content-based and content-free methods is proposed, which employs both short-term relevance feedback (STRF) and long-term relevance feedback (LTRF) as the means of user interaction. The STRF refers to iterative query-specific model learning during a retrieval session, and the LTRF is the estimation of a user history model from the past retrieval results approved by previous users. The framework is formulated based on the Bayes' theorem, in which the results from STRF and LTRF play the roles of refining the likelihood and the *a priori* information, respectively, and the images are ranked according to the *a posteriori* probability. Since the estimation of the user history model is based on the principle of collaborative filtering, the system is referred to as a collaborative Bayesian image retrieval (CLBIR) framework. To evaluate the effectiveness of the proposed framework, nearest neighbor CLBIR (NN-CLBIR) and support vector machine active learning CLBIR (SVMAL-CLBIR) were implemented. Experimental results showed the improvement over content-based methods in terms of both accuracy and ranking due to the integration in the proposed framework.

Index Terms— image retrieval, Bayesian framework

1. INTRODUCTION

Ever-lasting growth of multimedia information has been witnessed and experienced by human beings since the beginning of the information era. An immediate challenge resulting from the information explosion is how to intelligently manage and enjoy the multimedia databases. Content-based image retrieval (CBIR) has been intensively studied for more than a decade, yet still remaining a challenging topic [1]. Conventional CBIR systems exploiting global low-level features have proven effective to the extent of pre-attentive similarity due to the semantic gap. Noticing the critical role of human beings in recognizing semantic content in multimedia objects, relevance feedback (RF) was applied to CBIR. Modern techniques approach RF by approximating a function consistent with human visual perception [2–4], resulting in significant improvement. We refer to these RF techniques as short-term relevance feedback (STRF) as they are terminated once a user is satisfied by the results or gives up the query. On the other hand, we believe that a successful retrieval system should be capable of learning a history model of the vast majority of the users from the past retrieval results since they contain valuable semantic information which may improve the database-wide semantic indexing. We refer to the technique of learning a user history model as long-term relevance feedback (LTRF) because it can be a life-long process involving human computer interaction.

In this paper, we propose a new image retrieval strategy, in which the content-based and the content-free [5] methods are seamlessly integrated into a mathematically justifiable framework. User interaction is carried out through the combination of STRF and LTRF. We formulate the task based on the Bayes' theorem, in which the content-based similarity measure is considered as the likelihood evaluation which can be updated using STRF and the probability estimated using content-free approaches serves as the *a priori* information. The *a posteriori* probability is used to rank the images in the database. For the likelihood evaluation, we adopted both nearest-neighbor CBIR (NN-CBIR) and support vector machine active learning CBIR (SVMAL-CBIR). As for the content-independent component, we employed the MaxEnt-based CFIR. Numerical results demonstrated better performance than that of a simple content-based system with only STRF. In addition, even if there is no user history, the system can still function as the *a priori* distribution of the images is just uniform, in which case, however, the CFIR fails to work [6]. Since the *a priori* knowledge is extracted using a collaborative filtering technique, the proposed system is referred to as a collaborative Bayesian image retrieval (CLBIR) framework.

2. THE PROPOSED FRAMEWORK

Let a query be represented using a vector $\boldsymbol{x}_q$, where $\boldsymbol{x}_q \in \mathbb{R}^d$. The goal of the framework is to rank the candidate images using the the *a posteriori* probability $P(\omega|\boldsymbol{x}_q, \boldsymbol{I})$, where $\omega \in \boldsymbol{W}$ is the index of an image in a database, $\boldsymbol{W} = \{1, 2, \ldots, N\}$, N is the number of images, and $\boldsymbol{I}$ is the background information. According to the Bayes' theorem, the *a posteriori* probability of an image given a query can be written as

$$P(\omega|\boldsymbol{x}_q, \boldsymbol{I}) \propto p(\boldsymbol{x}_q|\omega, \boldsymbol{I})P(\omega|\boldsymbol{I}), \tag{1}$$

with the equality replaced by the proportionality due to the fact that the probability density function (PDF) of the observation $\boldsymbol{x}_q$ is a normalization constant given different ω. In the CLBIR framework, $\boldsymbol{I} = \{I_{q,1}, I_{q,2}, \ldots, I_{q,Q}\}$ is a set of the indexes of query images, where $I_{q,i} \in \boldsymbol{W}$, $i = 1, 2, \ldots, Q$, and Q is the number of query images. When $1 < Q \ll N$, $\boldsymbol{x}_q = \frac{1}{Q}\sum_{i=1}^{Q} \boldsymbol{x}_{q,i}$, where $\boldsymbol{x}_{q,i} \in \mathbb{R}^d$ is the feature vector of the query image $I_{q,i}$. According to the interpretation of $\boldsymbol{I}$, (1) can be simplified as

$$P(\omega|\boldsymbol{x}_q, \boldsymbol{I}) \propto p(\boldsymbol{x}_q|\omega)P(\omega|\boldsymbol{I}). \tag{2}$$

Based on (2), the information utilized for ranking candidate images consists of the similarity measure based on visual content and

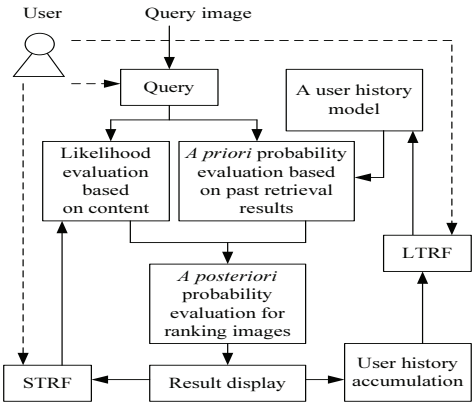


Fig. 1: The block diagram of the CLBIR framework.

past human judgement, i.e. past retrieval results approved by human users, corresponding to the likelihood evaluation and the *a priori* probability calculation. The block diagram of the proposed framework is illustrated in Fig. 1, in which the solid and dashed directed lines indicate the information flow and the human-controlled components, respectively. The STRF is employed to refine the content-based likelihood evaluation, whereas the LTRF is used to upgrade the statistical model characterizing the past retrieval history. Each retrieval session is composed of a number of STRF iterations and the LTRF is performed incrementally whenever a certain amount of new retrieval results are accumulated.

2.1. Content-based Likelihood Evaluation and STRF

2.1.1. Content-based Analysis Using NN-CBIR

The mechanism of NN-CBIR is to return the top K images on the list, which is ranked based on the similarity measure between the feature of the query and that of each of the candidate images, where $K \ll N$. In our framework, the L1-Norm is used as the distance function. For STRF, the query is refined using the method of query point movement. To calculate the likelihood, we employ the exponential function to convert the L1-Norm into a similarity function in order to approximate the likelihood in (2), i.e.

$$p(\boldsymbol{x}_q|\omega) = \frac{1}{A} e^{-|\boldsymbol{x}_q - \boldsymbol{x}_\omega|}, \tag{3}$$

where $A = \int e^{-|\boldsymbol{x}_q - \boldsymbol{x}_\omega|}$ is the normalization constant.

2.1.2. Content-based Analysis Using SVMAL-CBIR

SVM is a powerful tool for pattern recognition because it maximizes the minimum distance between the decision hyperplane and the training samples so as to minimize the generalization error. Given training samples $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_T, y_T)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ is the ground-truth label of $\boldsymbol{x}_i$, and $i \in \{1, 2, \ldots, T\}$, the optimal hyperplane can be represented as $f(\boldsymbol{x}) = \sum_{i=1}^{T} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$ where $K(\boldsymbol{x}_i, \boldsymbol{x})$ is the kernel function, α_i is the Lagrangian multiplier, and b is the bias. Due to the sparse sample problem of the RF in CBIR, the philosophy of active learning was introduced into the human-machine interaction, where the most informative images are shown to request user-provided labeling, resulting in the SVMAL-CBIR [4]. Since the output of an SVM with respect to a sample is the oriented distance from the sample to the hyperplane, the value could be either positive or negative. Therefore, the exponential function is employed again to convert the value of the discriminant function. When selecting radial basis functions as the kernel, we obtain

$$p(\boldsymbol{x}_q|\omega) = \frac{1}{A} e^{\sum_{i=1}^{T} \alpha_i y_i \exp \frac{-||\boldsymbol{x}_i - \boldsymbol{x}_q||^2}{\lambda} + b}, \tag{4}$$

where $A = \int e^{\sum_{i=1}^{T} \alpha_i y_i \exp \frac{-||\boldsymbol{x}_i - \boldsymbol{x}_q||^2}{\lambda} + b}$ is the normalization constant.

2.2. Content-free Analysis for Calculating the *A Priori* Probability and LTRF

The objective of the content-free analysis is to calculate the probability of a candidate image as a relevant one given a query, which is expressed as $P(\omega|\boldsymbol{I})$. First, we model each candidate image using a binary random variable Y_ν, where $\nu \in \boldsymbol{W}$. Y_ν has two states, with the state of $Y_\nu = 1$ indicating that the νth image is relevant and $Y_\nu = 0$ otherwise. Second, we model the query images using a set of binary random variables Y_J, where $J \subset \boldsymbol{W}$. Each element in Y_J also has two states, with the state of 1 indicating that the query image contains the semantic meaning in the user's information need and 0 otherwise. In what follows, we are only dealing with the probability of a variable with the state of being 1 without explicit representation, unless stated otherwise. Then, we estimate $P(Y_\nu|Y_J)$ using the MaxEnt approach [7], where $\nu \in \bar{J}$, which are finally used to approximate the $P(\omega|\boldsymbol{I})$.

The MaxEnt can be used to estimate the conditional probability $P(\boldsymbol{Y}_H|\boldsymbol{Y}_E)$, where $\boldsymbol{Y}_H = \{Y_{H,1}, Y_{H,2}, \ldots, Y_{H,M}\}$ and $\boldsymbol{Y}_E = \{Y_{E,1}, Y_{E,2}, \ldots, Y_{E,K}\}$ are referred to as the hidden set and the evidence set, and M and K are the respective sizes of $\boldsymbol{Y}_H$ and $\boldsymbol{Y}_E$. Since the goal is to calculate $P(Y_{H,h}|\boldsymbol{Y}_E)$, i.e. the probability of each hidden variable conditional on the evidence set, marginalization is needed, which can be rather computationally consuming given a large hidden set. Assuming the statistical independence across the hidden variables, the $P(Y_{H,h}|\boldsymbol{Y}_E)$ can be directly estimated by solving the following optimizations:

$$\max_{P_{h|E} \in [0,1]} - \sum_{y_{H,h}, y_E} \hat{P}(y_E) P_{h|E}^2 \tag{5}$$

subject to

$$\frac{\sum_{y_E} \hat{P}(y_E) P_{h|E} f_{E,k}}{\hat{P}(f_{E,k})} = \hat{P}(f_{H,h}|f_{E,k}),$$

where $k \in \{0, 1, \ldots, K\}$, $y_{H,h}, y_{E,k} \in \{0, 1\}$ represent the states of the hidden and evidence variables, and the $f_{H,h}$ and $f_{E,k}$ are hidden and evidence feature functions defined as $f_{H,h} = y_{H,h}$ and $f_{E,k} = y_{E,k}$ if $k \neq 1$ and $f_{E,k} = 1$ if $k = 0$, respectively. The optimization in (5) with respect to the conditional probabilities of the hidden variables can be carried out in parallel through matrix computation, which is referred to as the inverse probability method (IPM) [7]. The closed form solution to the optimization is

$$\mathbf{P}_{H|E} = \mathbf{PF}_{H|E} \times \mathbf{PF}_{E|E}^{-1} \times \mathbf{f}_E, \tag{6}$$

where

$$\mathbf{P}_{H|E} = [\hat{P}(X_{H,1}|\boldsymbol{x}_E), \hat{P}(X_{H,2}|\boldsymbol{x}_E), \ldots, \hat{P}(X_{H,M}|\boldsymbol{x}_E)]^T, \tag{7}$$

$$\mathbf{PF}_{H|E} \tag{8}$$
$$= \begin{pmatrix} \hat{P}(f_{H,1}) & \hat{P}(f_{H,1}|f_{E,1}) & \dots & \hat{P}(f_{H,1}|f_{E,K}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{P}(f_{H,M}) & \hat{P}(f_{H,M}|f_{E,1}) & \dots & \hat{P}(f_{H,M}|f_{E,K}) \end{pmatrix},$$

$$\mathbf{PF}_{E|E} \tag{9}$$
$$= \begin{pmatrix} 1 & 1 & \dots & 1 \\ \hat{P}(f_{E,1}) & 1 & \dots & \hat{P}(f_{E,1}|f_{E,K}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{P}(f_{E,K}) & \hat{P}(f_{E,K}|f_{E,1}) & \dots & 1 \end{pmatrix},$$

and

$$\mathbf{f}_E = [f_{E,0}(\boldsymbol{x}_E), f_{E,1}(\boldsymbol{x}_E), \dots, f_{E,K}(\boldsymbol{x}_E)]^T. \tag{10}$$

In the case under our consideration, $\boldsymbol{Y}_E = Y_J$ represents query images and $\boldsymbol{Y}_H$ is composed of the Y_ν's corresponding to the candidate images. Therefore, the conditional probabilities we are estimating can be written as

$$P(Y_\nu|Y_J) = \begin{cases} P(Y_{H,h}|\boldsymbol{Y}_E), & \nu \in \bar{J} \\ 1, & \nu \in J \end{cases}, \tag{11}$$

where $h \in \boldsymbol{W}/J$. With the above results we can approximate the $P(\omega|I)$ using

$$P(\omega|\boldsymbol{I}) = P(Y_\omega|Y_J) / \sum_{\nu=1}^{W} P(Y_\nu|Y_J). \tag{12}$$

When a certain amount of new retrieval results have been accumulated since the completion of the last update of the user history model, a new iteration of LTRF will be carried out using an incremental update procedure [6], by which the efficiency can be considerably raised, which is the amount of time needed for the refinement of the user history model.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

To guarantee the diversified image content, which is a typical situation of image retrieval in a large general domain, we randomly selected 200 classes from the COREL image collection, with 50 images in each class. The resultant 10000 images and the vendor-defined categories were used as the database and the ground truth for evaluating the performance. From the database, 10 queries were selected from each of the 200 classes, resulting in 2000 queries, each of which is composed of two different images. Under the query-by-example retrieval paradigm, the average of the features of the two images was used as the feature of each query. The queries were further divided into three mutually exclusive subsets, denoted by T_A, $T_{B,1}$, and $T_{B,2}$, where $|T_A| = 1000$, $|T_{B,1}| = 400$, and $|T_{B,2}| = 600$ are the respective sizes of the subsets. We employed global color histogram in Hue-Saturation-Value (HSV) space, color layout in YCbCr space and Gabor wavelet as low-level features. The experimental procedure is summarized as follows.

1) T_A was used when the user history model was not available, i.e. before LTRF happens. In such a case, only STRF is involved, and the NN-CLBIR and the SVMAL-CLBIR are essentially the same as the NN-CBIR and SVMAL-CBIR because the *a priori* distribution of the candidate images is uniform. The retrieval results corresponding to T_A were used to perform the initial LTRF.

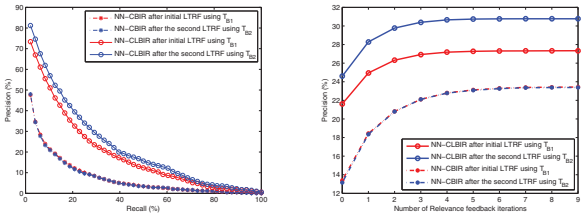
2) After the initial LTRF, the CLBIR systems are expected to present better performance thanks to the accumulated high-level knowledge characterized by the user history model, while the STRF still improves the results with respect to each specific query. $T_{B,1}$ was used to demonstrate the performance improvement.

3) During the operation of the CLBIR systems, the new retrieval results after a certain LTRF are gradually accumulated until the next LTRF is carried out. In our experiments, the retrieval results corresponding to $T_{B,1}$ were used to perform the second LTRF, i.e. an incremental update of the system. To show the effectiveness of the incremental update, the performance was evaluated using $T_{B,2}$. Since the query subsets are mutually exclusive, we guaranteed that the trained system using LTRF were tested based on previously unseen samples.

3.2. Numerical Results

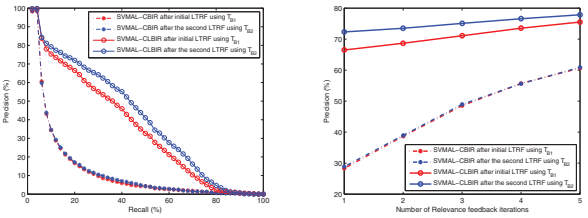
Shown in Fig. 2(b) is the comparison between NN-CBIR and NN-CLBIR in terms of the average precision P_{avg} as a function of the number of iterations of STRF, where the precision is defined as $P = \frac{N_C}{N_R}$, where N_C and N_R are the numbers of relevant images and retrieved images, respectively. We adopted $N_R = 48$ in this case. First, using query set $T_{B,1}$, the improvement due to the initial LTRF which was based on the retrieval results corresponding to T_A can be observed, showing the ability of the CLBIR systems to utilize the past retrieval results. Meanwhile, the improvement resulting from STRF can also be observed, which is shared by both CBIR and CLBIR systems. Second, after the second LTRF, the performance of NN-CLBIR using query set $T_{B,2}$ is further enhanced resulting from more accumulated knowledge through LTRF. Based on the same query set, the performance of NN-CBIR remains similar. To test the performance in terms of ranking ability, we employed the precision-versus-recall curve (PRC), where the recall is defined as $R = \frac{N_C}{N_G}$, where N_G is the number of images in the same semantic class as that of the query. The precision is averaged over all queries at each different recall value. The PRC after the initial retrieval was shown in Fig. 2(a). Higher precision value at a certain recall indicates more relevant images being ranked ahead of irrelevant ones, i.e. to reach the recall value, a smaller set of retrieved images has to be gone through. Based on this fact, the advantage of the integration of user history as high-level knowledge with the content analysis can be demonstrated based on the comparison in Fig. 2(a).

The comparison shown in Fig. 2(c) and Fig. 2(d) is for the same purpose of performance evaluation as that described above, and the difference lies with the approach to the content analysis for the likelihood computation, which is based on the output of the SVM employed for active learning-based STRF. In this case, we adopted $N_R = 20$ for the evaluation of precision as a function of the number of STRF iteration, and $N_G = 50$ for the evaluation of PRC. Since the initial retrieval is just random ranking, the precision was evaluated starting from the first STRF iteration. Still, we can observe the improvement resulting from the integration through the Bayesian framework.



(a) Comparison between NN-CBIR and NN-CLBIR.

(b) Comparison between NN-CBIR and NN-CLBIR.



(c) Comparison between SVMAL-CBIR and SVMAL-CLBIR.

(d) Comparison between SVMAL-CBIR and SVMAL-CLBIR.

Fig. 2: Objective evaluation on the performance improvement resulting from the proposed approach. a) and c) Comparison in terms of the PRC after the first retrieval iteration. b) and d) Comparison in terms of the precision as a function of the number of RF iterations.

3.3. Subjective Evaluation

An interface with the NN-CLBIR enabled has been implemented to demonstrate the effectiveness of the proposed framework in terms of performance improvement by the accumulation of user history. Illustrated in Fig. 3(a) and Fig. 3(b) are the top 20 retrieved images using NN-CLBIR. Shown in the upper figure is the result obtained using a system, whose *a priori* knowledge was extracted from 1000 past retrieval results, while in the lower one, the result is based on the *a priori* knowledge learned from 1400 past retrieval results. The query is selected from the semantic class of the theme soldier, and the last 4 images do not belong to this class in Fig. 3(a). Nonetheless, all of the top 20 images are relevant to the query in Fig. 3(b).

4. CONCLUSION

The integration of content-based and content-free methods for image retrieval is studied in this paper, which is formulated based on the Bayes's theorem. It employs STRF and LTRF for human machine interaction, which refines the query formulation and incrementally learns a user history model based on past retrieval results, respectively. The resulting framework can be considered as a CBIR system with memory. Moreover, it does not suffer from the cold start problem of CFIR. Two particular instances of the proposed framework has been implemented for experimental evaluation. Simulation results demonstrated the effectiveness of the combination of the content-based and content-free information, which include the improvement resulting from learning a user history model based on more accumulated knowledge, i.e. LTRF, and that by STRF during a retrieval session. Future work will be focused on seeking a more accurate approach to estimating the user history model.

(a) Based on the user history model trained using 1000 past retrieval results.

(b) Based on the user history model trained using 1400 past retrieval results.

Fig. 3: Retrieval results for subjective evaluation on the performance improvement resulting from more user history.

5. REFERENCES

[1] Ritendra Datta and et. al., "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1 – 5:60, April 2008.

[2] Paisarn Muneesawang and Ling Guan, "An interactive approach for cbir using a network of radial basis functions," *IEEE Transactions on Multimedia*, vol. 6, no. 5, pp. 703–716, Oct 2004.

[3] K. H. Yap and K. Wu, "A soft relevance framework in content-based image retrieval systems," vol. 15, no. 12, pp. 1557–1568, December 2005.

[4] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," *Proceedings of the ninth ACM international conference on Multimedia*, pp. 107–118, 2001.

[5] T. Kanade and S. Uchihashi, "User-powered 'content-free' approach to image retrieval," in *DLK*, March 2004, pp. 24 – 32.

[6] Rui Zhang and Ling Guan, "A new relevance feedback framework for content-free image retrieval," To appear in MMSP08.

[7] C. Zitnick, *Computing Conditional Probabilities in Large Domains by Maximizing Renyi's Quadratic Entropy*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, May 2003.