An Image Retrieval Scheme Using Multi-instance and Pseudo Image Concepts

Feng-Cheng Chang and Hsueh-Ming Hang*

Department of Electronics Engineering, National Chiao Tung University,
Hsinchu 300, Taiwan, R.O.C.,
fcchang.ee88g@nctu.edu.tw, hmhang@mail.nctu.edu.tw

Abstract. Content-based image search has long been considered a difficult task. Making correct conjectures on the user intention (perception) based on the query images is a critical step in the content-based search. One key concept in this paper is how we find the user preferred image characteristics from the multiple samples provided by the user. The second key concept is that when the user does not provide a sufficient number of samples, how we generate a set of consistent "pseudo images". The notion of image feature stability is thus introduced. In realizing the preceding concepts, an image search scheme is developed using the weighted low-level image features. At the end, quantitative simulation results are used to show the effectiveness of these concepts.

1 Introduction

The dramatically growing size of digital contents creates the demand for highly efficient multimedia content management. Each content-based image retrieval (CBIR) application requires a different set of configurations[1], including the selected image features and the processing architecture, to achieve the desired matching accuracy. There are no general guidelines in designing a good matching criterion; thus, many CBIR systems have been proposed to bridge the gap between image feature space and human semantics.

In this paper, we will focus on the content-based image retrieval (CBIR) methods. In sec. 2, we briefly discuss the concept of multiple instances and common problems when using this technique. Based on a few assumptions, we propose a straightforward yet effective method that incorporates multiple samples and image multi-scale property in estimating user intention in sec. 3. Then, the subjective and objective performance of the proposed scheme is shown in sec. 4. At the end, we conclude this presentation with sec. 5.

* This work is partially supported by the Lee & MTI Center for Networking Research at National Chiao Tung University and by National Science Council (Taiwan, ROC) under Grant NSC 91-2219-E-009-041

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3331, pp. 157–164, 2004.

2 Motivations

In a typical Query-by-Example (QBE) CBIR system with relevance feedback function, it analyzes the user query images and/or relevant feedback images to derive the search parameters. The search parameters are often defined in terms of the image features pre-chosen in the system. Then the system searches the database and returns a list of the top-N similar images for further relevance feedback. This process can be repeated and hopefully it will eventually produce the satisfactory results to that particular user and query. In such a system, multiple samples (query and feedback) help the system to make a better "guess" on the user intention.

The problem is how one utilizes multiple image features and multiple query instances (images) to derive the proper search parameters. Multiple features and multiple instances represent two different aspects. The former is how we describe an image in an application; the latter is how we guess the user intention using the given instances. There exist many proposals on combining multiple features for image search such as [2]. Methods of combining multiple instances are usually considered as a part of a relevance feedback function. There are several existing CBIR proposals containing relevance feedback such as MARS[3,4] and iPURE[5].

In our previous project, we developed an MPEG-7 test bed [6] and thus have used it to examine several low-level MPEG-7 features. We observed that subjectively similar pictures tends to be close (near) in one or more feature spaces. Another observation is that a low-level feature often has (somewhat) different values when it is extracted from the same picture with different spatial resolutions and/or picture quality (SNR scalability). Our investigation finds that people often design a QBE system with feedback under the assumption that a sufficient number of query instances or feedback iterations can be provided by the user. However, this assumption is not always true in a real-world application[7]. Often, the sample size is very small (one to three) and the information contained in various samples may not be all consistent. Based on our observations, we are motivated to develop a distance-based user perception estimation algorithm, which tries to produce a correct conjecture on the user intention based on the small number of samples (instances) provided by the user.

3 Weighting on Low-Level Features

In the following discussions, we focus on a statistical approach that combines multiple low-level features together to form a "good" metric for retrieving "similar" images. We first describe the feature weights produced by multiple instances (query set) in sec. 3.1. Then, the approach of generating pseudo images using multiple (resolution or SNR) scales is described in sec. 3.2. In sec. 3.3, we present a CBIR architecture that uses the multi-instance and pseudo image concepts. It solves the feature space normalization problem, and reduces the impact of insufficient user feedback information.

3.1 User Perception Estimation

There are several ways to combine different low-level features. Here we use a straightforward one: weighted sum of feature distances. This method is simple because we use the ready-to-use distance functions, and the user perception is expressed by a weighting vector. Note that the weighting vector is derived from the multiple instances provided by the user.

Similar to many other image retrieval schemes, we assume the following conditions are satisfied:

- All the basic feature distance metrics are bounded.
- Two perceptually similar images have a small distance in at least one feature space.
- Low-level features are locally inferable[8]. That is, if the feature values of two images are fairly close, then the two images are perceptually similar.

In addition to the above assumptions, we add another conjecture: if two images have a large distance value in a specific feature space, we cannot determine the perceptual similarity of them based merely on this feature. Note that this feature space is simply irrelevant to our perception. It does not necessarily decide dissimilarity in perception.

Different from several well-known CBIR systems, our system does not rely on *a priori* feature distributions. These distributions may help to optimize inter-feature normalization, as in [3], to produce better performance in accuracy. However, they often introduce overheads and degrade system performance in speed. Even if feature distributions are available, they may not lead to appropriate normalization. Thus, we try to design our method to be independent of feature distributions as shown below. The need of normalization is eliminated because of the way we define distance function.

In summary, our feature weighting and combination principle is: *given two user-input query images, if they are farther apart in a certain feature space, this feature is less important in deciding the perceptual similarity for this particular query.* Suppose we have a query image set with n samples, $Q = \{q_i \mid i = 1..n\}$, and an available basic feature set $F = \{F_j \mid j = 1..m\}$. Let f_{ij} denotes the feature F_j value for image q_i. The normalized distance function for feature F_j is $d_j(f_{1j}, f_{2j}) = n_j * D_j(f_{1j}, f_{2j})$, where $D_j(f_{1j}, f_{2j})$ is the designated distance function for F_j, and n_j is the normalization factor for F_j, which sets the normalized value $d_j(f_{1j}, f_{2j})$ in the range of $[0, 1]$. Though n_j is an *a priori* information, we will see that it can be safely discarded at the end of this section.

We next define the feature difference between image i and all the other images in Q for feature F_j as follows:

$$\mathit{diff}_{ij} = \mu_{ij} + \sigma_{ij},$$

where

$$\mu_{ij} = \frac{1}{n-1} \sum_{k=1, k \neq i}^{n} d_j(f_{ij}, f_{kj})$$

$$\sigma_{ij}^2 = \frac{1}{n-1} \sum_{k=1,k\neq i}^{n} (d_j(f_{ij}, f_{kj}))^2 - \mu_{ij}^2.$$

The extra term (standard deviation) is added into the difference measure because experiments indicate that an "inconsistent" feature (large standard deviation) is less important. Then we express the scatter factor as the maximum difference in this feature space: $s_j = \max_{\forall i} \mathit{diff}_{ij}$. The scatter factors can be considered as the importance indicator of that feature. Based on the previously mentioned rule, we give less perception weight to a more scattered feature (F_j):

$$w_j = (s_j * \sum_{k=1}^{m} \frac{1}{s_k})^{-1}.$$

The distance function combining m features is then defined as

$$D(q_1, q_2) = \sum_{j=1}^{m} w_j * d_j(f_{1j}, f_{2j}).$$

Finally, the distance function between image I and n query instances (Q) is defined by

$$D(I, Q) = \min_{i=1..n} D(I, q_i).$$

Note that the normalization factor n_j is canceled in every $w_j * d_j(f_{1j}, f_{2j})$ term. This implies that we can safely ignore the distance normalization problem as long as all the feature metrics are bounded.

3.2 Pseudo Query Images

In case that the number of query images is too small, we use the multi-scale technique to create pseudo query images. The term "scale" here refers to either the spatial resolution or the SNR quality. It is based on the conjecture that the down-sampled or noise-added images are subjectively similar to the original version. We also observe that a low-level feature may have different values at different scales (in spatial and in SNR).

An unstable (sensitive) feature tends to yield a large distance value when the distance is computed based on different scales of the same image. The quantitative difference in stability can be measured by the scatter factor s_j defined in sec. 3.1. Therefore, we adopt another principle: *we have a higher matching confidence (more weight) on the distance metric associated with a stable feature.* Now, we can include the stability estimation into the perception estimation by adding these pseudo images to the query set. The combined procedure thus puts less weight on more scattered features, which may be due to either perceptual irrelevance or feature instability.

3.3 Architecture

The proposed CBIR query system architecture is summarized in Fig. 1. The original query (input) images are processed to produce pseudo-images. Together they form the query set. The query set is fed into the user perception analysis process to estimate the weighting factors. Then, the query set and the weighting factors are sent to the image matching process to compute image similarity. At the end, the process generates the top-N list.

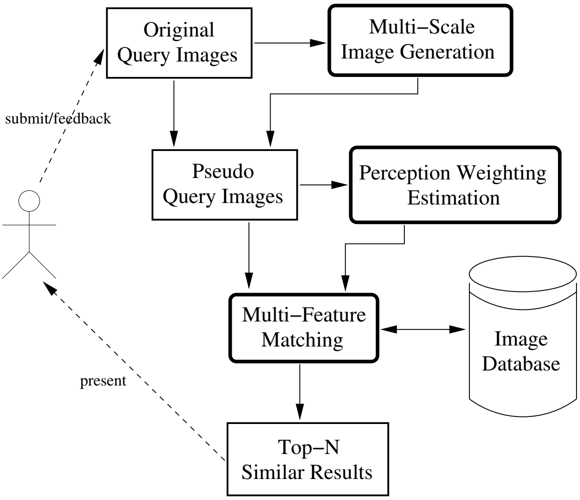


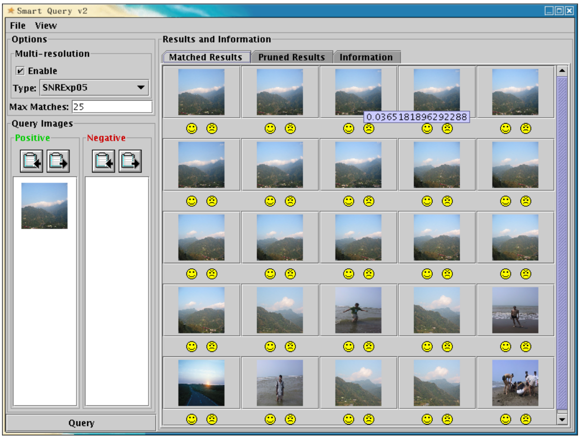
Fig. 1. Proposed perception estimation and query system

4 Experiments and Discussions

In this section, we examine our design using both subjective and objective measures. The screenshot shown on Fig. 2 is an application program running on our MPEG-7 test bed [6]. Three image global features defined by MPEG-7 are used. They are scalable color, color layout, and edge histogram. The query images are displayed on left panel. The right panel shows the top-25 query results.

4.1 ANMRR

We adopt the *Average Normalized Modified Retrieval Rank* (ANMRR)[9] metric in measuring the accuracy of our method. The ANMRR is used in the MPEG-7 standardization process to quantitatively compare the retrieval accuracy of different competing visual descriptors. For a query image, this measurement favors a matched ground-truth result and penalizes a missing ground-truth or



☑ Enable

Type: SNRExp05

Max Matches: 25

Fig. 2. Subjective results

a non-ground-truth result. We briefly describe the formula of ANMRR in the following paragraphs. Details can be found in [9][10, pp.183-184].

For a query q with a ground-truth size of $NG(q)$, we define $rank(k)$ as the rank of the kth ground-truth image on the top-N result list. Then,

$$Rank(k) = \begin{cases} rank(k) & \text{if } rank(k) \leq K(q) \\ 1.25 \cdot K(q) & \text{if } rank(k) > K(q) \end{cases}$$

$$K(q) = \min\{4 \cdot NG(q), 2 \cdot \max[NG(q), \forall q]\}.$$

The average retrieval rank is then computed and normalized with respect to the ground-truth set to yield the *Normalized Modified Retrieval Rank* (NMRR):

$$NMRR(q) = \frac{\frac{1}{NG(q)} \sum_{k=1}^{NG(q)} Rank(k) - 0.5 \cdot [1 + NG(q)]}{1.25 \cdot K(q) - 0.5 \cdot [1 + NG(q)]}.$$

The range of $NMRR(q)$ is $[0, 1]$. The value 0 indicates a perfect match that all the ground-truth pictures are included in the top-rank list. On the other hand, the value 1 means no match. Finally, we have the *Average Normalized Modified Retrieval Rank* (ANMRR):

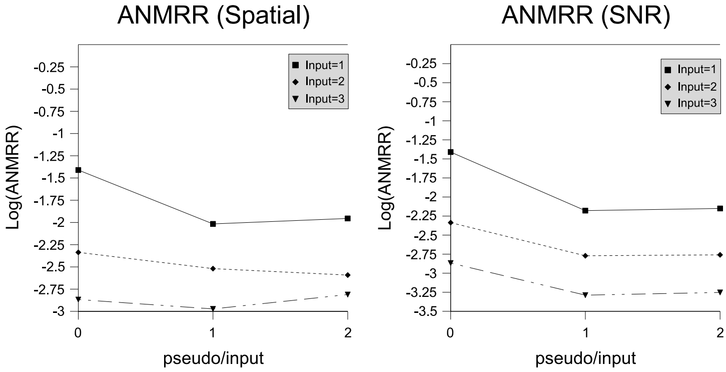
$$ANMRR = \frac{1}{NQ} \sum_{q=1}^{NQ} NMRR(q),$$

where NQ is the number of queries.

4.2 Experiments and Results

Our test images are pure scenic images. We collect 38 sets of scenic images as the ground truth. Each set of ground-truth images is taken on the same spot with slightly different camera pan and tilt angles. The size of a ground-truth set varies from 4 to 10. With additional randomly selected images, the database contains 1050 images in total.

Two multi-scale schemes are simulated: spatial and SNR. The spatial scaling factor (both width and height) for each down-sampled image is defined as follows: the n-th scale factor is $\alpha^n (\alpha = 0.7)$. In Fig. 3(a), we examine the effect of different pseudo/input image ratios. Under the same pseudo/input ratio, the more the input images (user provided), the better the query accuracy. For the same number of input images, pseudo images can improve the accuracy, especially when the input images is one or two. However, when input (query) images are higher in number, the addition of pseudo images may lower the matching accuracy. Fig. 3(b) shows the results of using SNR-scaled pseudo images. The noisy versions (pseudo images) are generated by applying JPEG compression with a quality factor of $\beta^n (\beta = 0.4)$ for the n-th scaled version. The SNR results are similar to those of spatial-scaled, with the exception that the average ANMRR is better in SNR multi-resolution approach.



(a) (b)

Fig. 3. Simulation results

5 Conclusion

In this paper, the problem associated with multi-instance image retrieval is investigated. The main contributions of this paper are (1) propose a distance-based method to estimate user perceptions based on the given multiple instances, and (2) generate consistent pseudo images when the query set is too small. The first concept is realized by analyzing the scattering of the query instances in feature space. Our conjecture is that a scattered feature implies less importance in deciding the perceptual similarity. The second concept is realized through the notion of feature stability. Our conjecture is that a stable image feature (for a particular image) has similar numerical values (small scatter factor) at different spatial or SNR scales of the same image. Therefore, pseudo images are created by scaling the original image at various spatial and SNR resolutions.

All the preceding concepts can be integrated into one algorithm using the same basic structure – adjusting the weights of features. We examined the performance of our scheme using MPEG ANMRR. Simulations show that multiple instances are helpful in achieving better query accuracy. In the case that the user input set is small, the synthesized pseudo images also improve the results in most cases. There are several parameters and/or distance measures can be further fine-tuned to produce better results.

References

1. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
2. Jeong, S., Kim, K., Chun, B., Lee, J., Bae, Y.J.: An effective method for combining multiple features of image retrieval. In: IEEE TENCON. Volume 2. (1999) 982–985
3. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol. **8** (1998) 644–655
4. Rui, Y., Huang, T.S., Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS. In: Proc. IEEE Int. Conf. Image Processing. (1997) 815–818
5. Aggarwal, G., V., A.T., Ghosal, S.: An image retrieval system with automatic query modification. IEEE Trans. Multimedia **4** (2002) 201–214
6. Chang, F.C., Hang, H.M., Huang, H.C.: Research friendly MPEG-7 software testbed. In: Image and Video Communication and Processing Conf., Santa Clara, USA (2003) 890–901
7. T.V., A., Jain, N., Ghosal, S.: Improving image retrieval performance with negative relevance feedback. In: ICASSP. (2001) 1637–1640
8. Zhang, C., Chen, T.: An active learning framework for content-based information retrieval. IEEE Trans. Multimedia **4** (2002) 260–268
9. Committe, M., ed.: Subjective Evaluation of the MPEG-7 Retrieval Accuracy Measure (ANMRR). ISO/IEC JTC1/SC29/WG11, M6029, MPEG Committee (2000)
10. Manjunath, B.S., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7. John Wiley & Sons Ltd., Baffins Lane, Chichester, West Sussex PO19 1UD, England (2002)