A Relevance Feedback Image Retrieval Scheme Using Multi-Instance and Pseudo Image Concepts

Feng-Cheng Chang and Hsueh-Ming Hang[a]

[a]Department of Electronics Engineering, National Chiao Tung University, Taiwan, R.O.C.

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 positive 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. The third key concept is how we use negative images as pruning criterion. 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.

Keywords: image retrieval, perception weighting

1. INTRODUCTION

Due to the increasing popularity of digital capturing devices such as digital camera, the dramatically large size of digital contents demands for highly efficient multimedia content management. For a particular application, a content-based image retrieval (CBIR) system often has a distinct set of configurations[1] including the selected image features and the processing architecture, in order to achieve the desired matching accuracy. A known approach for constructing a human-satisfactory CBIR system is to incorporate semantic related features for matching. However, there are no general guidelines in designing or acquiring these features; thus, many CBIR systems have been proposed to bridge the gap between image feature space and human semantics by relevance feedback. In this paper, we will focus on the content-based image retrieval (CBIR) methods. In Sect. 2, we briefly discuss the concept of multiple instances and the problems in using this technique. Based on a few assumptions, we propose a straightforward yet effective method that incorporates multiple samples and image multi-scale property for estimating user intention in Sect. 3. Then, the subjective and objective performance of the proposed scheme is shown in Sect. 4. At the end, we conclude this presentation with Sect. 5.

2. MOTIVATION

Researches on CBIR topics show that semantic related features are critical in boosting the query accuracy. These high-level features (semantic information) can be either extracted at the stage of content analysis, or acquired at run-time. The former is used to provide semantical content descriptions and often requires sophisticated or manual feature extraction processes, such as object segmentation or textual descriptions. The latter is often used to capture the user preferences, and often tries to derive the information from run-time input sources, for example, user feedbacks. We are interested in estimating user perceptions on-the-fly. To simplify the simulation and clarify the effectiveness of the proposed estimation scheme, the goal of our system is to use only low-level features for automatically high-volume feature extraction and matching, and to derive user perceptions from available input images as precisely as possible. We also consider a simple user interface for relevance feedback, and thus the application only asks users to label positive or negative images in a free way. (No specific number of feedback images is required.)

Further author information: (Send correspondence to Feng-Cheng Chang)
Feng-Cheng Chang: E-mail: fcchang.ee88g@nctu.edu.tw
Hsueh-Ming Hang: E-mail: hmhang@mail.nctu.edu.tw

Storage and Retrieval Methods and Applications for Multimedia 2005, edited by Rainer W. Lienhart, Noboru Babaguchi, Edward Y. Chang, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5682 © 2005 SPIE and IS&T · 0277-786X/05/$15

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 using Borda counts.[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 testbed[6] and thus have used it to examine several low-level MPEG-7 features. We observed that subjectively similar pictures tend 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 the 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 make a correct conjecture on the user intention based on the small number of samples (instances) provided by the user.

3. PROPOSED WEIGHTING METHOD

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 Sect. 3.1. When a user selects an image as a negative example, we use the method described in Sect. 3.2 to prune irrelevant results. Then, the approach of generating pseudo images using multiple (spatial or SNR) scales is described in Sect. 3.3. In Sect. 3.4, we propose 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 supplied 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. And the user perception is expressed by a weighting vector. Note that the weighting vector is to be 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 all 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 MARS,[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. More importantly, user perceptions do not necessarily match the feature distributions in the database. 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} denote the value of feature F_j 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 (distance) 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 second 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 number* as the maximum difference in this feature space: $s_j = \max_{\forall i} \mathit{diff}_{ij}$. The scatter numbers may be interpreted as an importance indicator of that feature. If some members in this group are far away from the rest, the scatter number of the entire group is large because the maximum distance is chosen. Based on the previously described principle, we give less *perception weight* to a more scattered feature (F_j):

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

The distance function (of two images, q_1 and q_2) 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. Irrelevant Images

In this section, we will describe how to use irrelevant (negative feedback) images to improve the query accuracy. For a typical QBE search, a user provides a non-empty set of relevant (positive feedback) images. Suppose we also ask the user to select negative (irrelevant) images. In our proposed scheme, negative images are not included in computing the perceptual weights. This is due to the following observations. (1) The human perception of similarity and dissimilarity may not be (linearly) additive. (2) When an image is considered dissimilar to the query one, we do not know which features (one or many) dominate in producing the perceptual dissimilarity.

So we use negative images in the following way: they create "holes" in the feature space. That is, the database images located inside the pruning radius and close to the negative images are removed from the top-N (similar) list. Essentially, we conduct a pruning process for removing positively correlated images based on the given negative image(s). Let Q_p and Q_n are the positive and the negative image sets respectively. A *pruning radius* associated with a negative image $g_i \in Q_n$ is specified by $r_p(g_i) = D(g_i, Q_p)$. An image I_r is thus removed from the top-N list if I_r is located in a pruning region:

$$\exists g_i \in Q_n \text{ satisfies } \begin{cases} D(I_r, g_i) < r_p(g_i) \\ D(I_r, Q_p) > D(I_r, g_i) \end{cases} .$$

There are two conditions given in the preceding equation. An intuitive explanation to the second condition is that if an image is closer to Q_n than Q_p, it is excluded from the top-N list. But since the negative feedback sample set is small and incomplete, we do not want to exclude the images that are a bit far away from both sets but are slightly closer to a negative sample. Therefore, the first condition gives a maximum pruning radius. Thus, our pruning operation starts from the highest priority item on the top-N list. If an image is closer to Q_n than Q_p and is located inside the pruning radius, it is excluded from the top-N list.

3.3. 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 often have somewhat different values at different scales (in spatial and in SNR).

An unstable (sensitive) feature in our definition yields a large distance value among the derived images from the original at various scales. The measure of instability is again specified by the scatter number s_j defined in Sect. 3.1. Stable features often represent the most noticeable features of an image and they in term are often the features that the inquiring users desire. Therefore, we come up with another principle: *We give the stable features of a query image more confidence (more weight) in searching for its similar images.* Thus, we include these pseudo images into 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. Two possible pseudo image generation methods are described later in Sect. 4. Although the pseudo image concept is motivated by the lack of input images, we will see that it also improves accuracy when the number of input images is one or two. Hence, the second principle is justified mostly by observations and experiments.

3.4. Architecture

The proposed CBIR query system architecture is summarized by Fig. 1. The original positive query (input) images are used to generate 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 passed to the image matching process to compute image similarity. A tentative matching list is thus produced. Then, the pruning process based on the supplied negative images is applied to the tentative matching list and some "irrelevant" images may be removed. At the end, we receive the final top-N list.

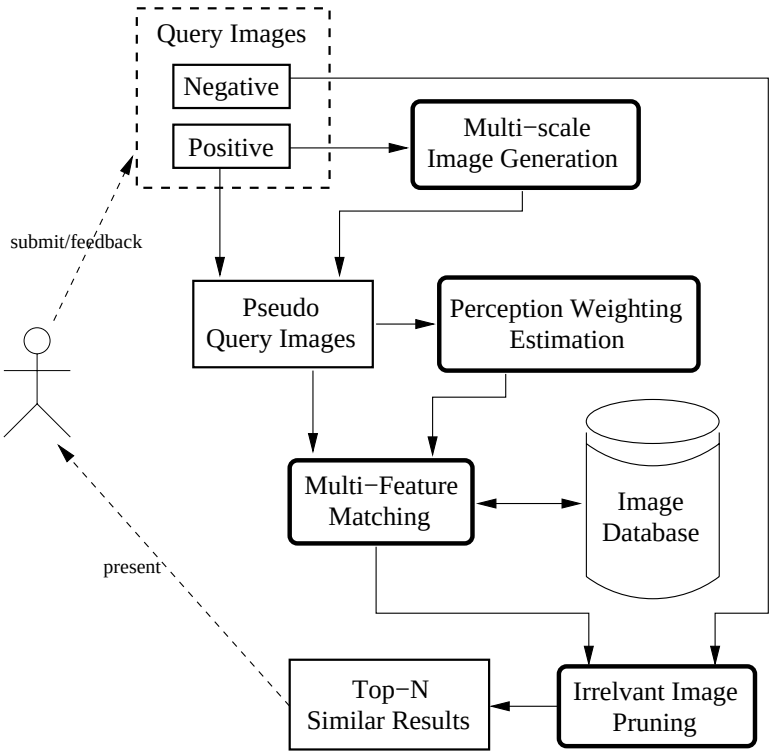

Figure 1. Proposed perception estimation and query system

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 testbed.[6] Three image global features defined by MPEG-7[9] are adopted. They are scalable color, color layout, and edge histogram. The query images are displayed on the left panel. The right panel shows the top-25 query results. The happy/unhappy buttons under each result image represent the positive/negative feedback options, respectively. Once a button is pressed, its associated image is added to the positive or negative list on the left panel as the next-iteration input image.

4.1. ANMRR

Many CBIR researches use precision and recall analysis to rate the accuracy of a system. In fact, these two rating methods represent two different viewpoints. The former one is the ratio between the number of the retrieved relevant images and the number of the total retrieved images. The latter is the ratio between the number of the retrieved relevant images and the number of the pre-defined relevant (so-called ground truth) images. These two rates are influenced by the chosen size of the top-N list. To find a more objective measure, we adopt the *Average Normalized Modified Retrieval Rank* (ANMRR)[10] metric. The ANMRR is used in the MPEG-7 standardization process to quantitatively compare the retrieval accuracy of competing visual descriptors. This metric is a modified combination of precision and recall metrics, and is a normalized index to rate the overall query accuracy. For a query image, this measurement favors a matched ground-truth result and penalizes a missing ground-truth or a non-ground-truth item on the top-N list. We briefly describe the formula of ANMRR in the following paragraphs. Details can be found in the references.[10, 11]

For a query image 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,

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

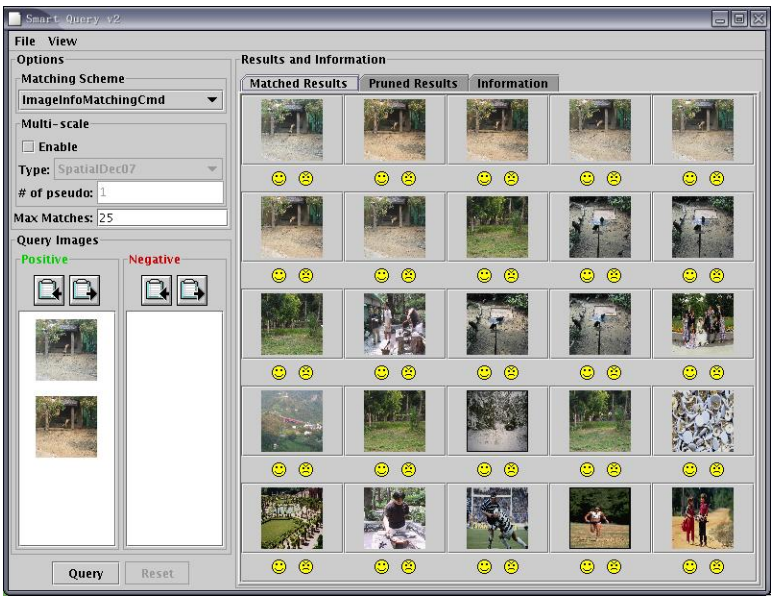

Figure 2. Subjective results

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

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

In our previous work,[12] we have conducted a preliminary experiment to evaluate the proposed method against a 1050-image database. The results show that the multi-instance user perception weighting method is promising, and the pruning concept always improves the query accuracy in our method; also, the pseudo-image concept improves the accuracy in many cases.

In this paper, we extend the evaluation process to a much larger scale. The database consists of 18433 images including 256 test (ground-truth) images, 194 people (party) photos, 200 flower pictures, 200 undersea pictures, 200 outdoor scenery pictures, and 17383 images from the Corel gallery.

We collect 38 sets of outdoor 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 by hand. The size of a ground-truth set varies from 4 to 10. Images in each set are perceptually similar. However, by examining the low-level features, we observe that the feature values can be quite different. There are several possible causes. One is that the

hand-taken photos suffer from shaking, which produces rotated or blurred images. The other is that different shots have slightly different focus and shutter speed. Another is that photos with shooting angle variation may have different background lighting, which may change the white-balance of each picture.

Our experiments simulate a typical image query scenario. A user first chooses one or more "similar" input images to start a query. The matching process returns an ordered list of results, and we call it the positive-only query result. If the result is not perfect; that is, not all ground-truth images occupy the highest ranks, or simply $NMRR \neq 0$, then the highest ranked non-ground-truth image is assigned as the negative feedback item. Then, we repeat the query process with both positive and negative images and produce the positive-and-negative query result. If the positive-only result is perfect, both $NMRR_{positive-only}$ and $NMRR_{positive-and-negative}$ are set to zero. Since the smallest ground truth set has only four images, we simulate the conditions of one to three positive images per query. All possible combinations of images in all ground truth sets are tested to derive the $ANMRR$ values.

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 (for the n-th pseudo image) is $\alpha - 0.1(n-1)$, where $n = 1, 2$. We perform experiments at $\alpha = 0.9, 0.8, 0.7, 0.6, 0.5$ to look for the best parameter values that lead to the best ANMRR. The SNR-scaled images are generated by applying JPEG compression with a quality factor of $\beta - 0.1(n-1)$ for the n-th scaled version. The test values are $\beta = 0.7, 0.6, 0.5, 0.4, 0.3$.

To examine the effectiveness of our method, we simulate another two weighting schemes under the same assumptions. The first scheme is a variation derived from the MARS system. In this scheme, distance metric $d_j(f_1, f_2)$ for each feature F_j is normalized as follows:

$$d'_j(f_1, f_2) = \frac{D_j(f_1, f_2) - \mu_j}{3\sigma_j},$$

where μ_j and σ_j^2 are the mean and variance of the distances of F_j in the database. This step ensures that about 99% of the distance values are within the range of $[-1, +1]$. The second parameter-shifting step guarantees that these 99% values are within $[0, 1]$:

$$d''_j(f_1, f_2) = \frac{d'_j(f_1, f_2) + 1}{2}.$$

The final step clamps all calculated distance values between zero and one.

The original MARS system adopts a 5-level relevance feedback. To make it comparable with our simulation environment, we reduce the relevance feedback levels to three: relevant ($Score_l = +1$), no opinion ($Score_l = 0$), and not relevant ($Score_l = -1$). The weighting process is similar to that in the original MARS. Assume the overall query result list is RT, and the result list of feature F_j is RT_j. To calculate the weight w_j, we first initialize $W_j = 0$, and then update W_j as follows:

$$W_j = W_j + Score_l, \text{ for each item } l \text{ which appears in both } RT \text{ and } RT_j.$$

After all W_j have been updated, we compute the weighting factor for each feature F_j as

$$w_j = \frac{W_j}{\sum_{\forall j} W_j}.$$

A final remark about this MARS-like scheme is the RT list. According to the original proposal,[3] the procedure is an iterative estimation process to approach the "optimal" RT. The original proposal selects $P_{fd} = 3$ as the maximum number of iterations and shows good convergence in general. In our simulation, we set $P_{fd} = 5$. This is Scheme A.

The second scheme we simulate has the same basic structure as our proposed scheme in Sect. 3. However, it adopts the same distance normalization as in MARS. This is Scheme B. We simulate this scheme for two reasons. One is to compare with a MARS-like scheme to see the effects of different weighting estimation procedures. The other is to compare it with our scheme to see the effects of different distance normalization methods. Our scheme is labeled as Scheme C.

	Scheme A: MARS-like	Scheme B: Gaussian-Normalized	Scheme C: Max-val Normalized
Input/Query=1			
Pseudo/Input=0	**-2.23**	**-2.23**	-1.38
Pseudo/Input=1	**-2.22**	-2.18	-1.94
Pseudo/Input=2	**-2.19**	-2.15	-1.93
Input/Query=2			
Pseudo/Input=0	-2.40	**-2.45**	-2.30
Pseudo/Input=1	-2.40	-2.45	**-2.51**
Pseudo/Input=2	-2.33	-2.37	**-2.50**
Input/Query=3			
Pseudo/Input=0	-2.40	-2.48	**-2.83**
Pseudo/Input=1	-2.41	-2.48	**-2.92**
Pseudo/Input=2	-2.33	-2.38	**-2.92**

Table 1. Best $log(ANMRR)$ of spatial-scaled pseudo images (positive-only)

4.3. Simulation Results

The simulation results are shown in Tables 1, 2, 3, and 4. The bold-faced numbers are the best accuracy among all tests with the same query parameters, and the underlined numbers are the worst ones. The row of *Input/Query* means the number of positive input images selected by the user in a query. The row of "Pseudo/Input" means the number of *pseudo images* created from each (user selected) input image. The first column is the MARS-like scheme, and the second and third columns are our schemes with different normalization formulas. To see more clearly the differences among various methods and parameters, the ANMRR values are listed in log scale. In the following paragraphs, we will examine these results and discuss the performance of various methods.

4.3.1. Observations on Positive-only Query Results

First we examine the results of positive-only spatial-scaled experiments (Table 1). For each method, multiple input images (all the cases where Pseudo/Input=0) improve the query accuracy. This shows that more "positive" information would result in better query precision, regardless which scheme is in use. Next, we examine the effect of pseudo images. Our scheme with one pseudo image has the best accuracy in all Input/Query cases. However, the pseudo images do not improve the other two methods as much. Even worse, more pseudo images would degrade the query accuracy. The simulation results also show that increasing pseudo images does not always improve accuracy. Under our current scheme, one pseudo image per input image is the best.

For each query parameter set (Input/Query and Pseudo/Input), we compare results of different estimation methods. Comparing the MARS-like method (Scheme A) with the Gaussian-normalized method (Scheme B), we observe that when input images are few, the MARS-like scheme wins. In contrast, Scheme B wins when more input images are provided. Since these two methods use the same distance normalization procedure, the difference comes from the weight computing procedures. When few images are available for estimation, iterative training would provide a better guess on the user perception. When more images are provided by a user, ranking-list-based MARS-like scheme does not provide as precise guess as distance-based Gaussian-normalized scheme. Comparing the Gaussian-normalized scheme (Scheme B) with our scheme (Scheme C), the former wins when input images are few and loses when more images are provided. The two methods use the same distance definition and the estimation procedure, so the difference comes from the distance normalization procedures. It is reasonable that the Gaussian-normalized scheme wins for few inputs cases, because the distance metrics are optimized according to the data distribution. This implicitly provides clustering information of the database, and thus produces better results than our method. However, feature distributions in a database may not be the same as the distance distribution viewed from the user perception of a particular query. This may explain why our method wins when more input images are provided. We estimate the user perception (intention) only based on the user provided information (not the database).

	Scheme A: MARS-like	Scheme B: Gaussian-Normalized	Scheme C: Max-val Normalized
Input/Query=1			
Pseudo/Input=0	**-2.23**	**-2.23**	-1.38
Pseudo/Input=1	**-2.19**	-2.18	-1.95
Pseudo/Input=2	**-2.18**	**-2.18**	-1.99
Input/Query=2			
Pseudo/Input=0	-2.40	**-2.45**	-2.30
Pseudo/Input=1	-2.45	-2.49	**-2.52**
Pseudo/Input=2	-2.47	-2.49	**-2.52**
Input/Query=3			
Pseudo/Input=0	-2.40	-2.48	**-2.83**
Pseudo/Input=1	-2.54	-2.63	**-3.07**
Pseudo/Input=2	-2.63	-2.71	**-3.10**

Table 2. Best $log(ANMRR)$ of SNR-scaled pseudo images (positive-only)

We apply the same analysis to the SNR-scaled case (Table 2), and similar conclusions are observed. However, there are two noticeable differences. The first one is that in several test cases, Pseudo/Input=2 outperforms Pseudo/Input=1. The second is that in most cases, the SNR-scaled pseudo images outperforms the spatial-scaled ones.

4.3.2. Observations on Positive-and-Negative Query Results

Next, we look into the Positive-and-Negative Query cases. Similar to what we did in Sect. 4.3.1, we first examine the simulation results of positive-and-negative feedback with spatial-scaled pseudo images, (Table 3). For all schemes, multiple input images improve the accuracy. Effects of pseudo images are similar to that of the positive-only results. Our method seems to be able to utilize pseudo images better for improving the accuracy. For the other two schemes, pseudo images do not provide significant improvements. The ANMRR values show that Pseudo/Input=1 cases give the most significant improvement. Additional pseudo images offer much less improvement if any.

Comparing Scheme A (MARS-like scheme) with Scheme B (Gaussian-normalized scheme), we found that the Gaussian-normalized scheme wins in most cases. Our explanation is that in our proposed procedure, the negative feedback does not participate in weights estimation. Since the negative instances may be too diverse to be useful in weights estimation, their role are more appropriate when used in pruning. The simulation results seems to prove this concept. Comparing Scheme C (our scheme) with Scheme B (Gaussian-normalized scheme), the results show that ours wins when sufficient input images are available. The ANMRR values show that the best accuracy is the Input/Query=3 case in our scheme. The reason is that the pruning distance relies on the estimated distance function. Thus, the more precise distance function would lead to a lower "mis-pruning" probability.

The ANMRR values shown in Table 4 for SNR-scaled pseudo images lead to similar conclusions as before. First, multiple input instances improve query accuracy. Second, our method benefits more from the pseudo images. Third, the Gaussian-normalized scheme (Scheme B) wins in almost all cases when comparing to the MARS-like scheme (Scheme A). Fourth, our method (Scheme C) performs better than the Gaussian-normalized when more input images are available. Finally, our method has a significant performance improvement at Input/Query=3, which indicates a good potential of our approach for even more input images.

4.3.3. Observations on Different Feedback Schemes

In Sect. 4.3.1 and Sect. 4.3.2, we discuss the effects of different weights estimation methods in each specific scheme. In this section, we will discuss the general effect of negative instances and the generation of pseudo images.

	Scheme A: MARS-like	Scheme B: Gaussian-Normalized	Scheme C: Max-val Normalized
Input/Query=1			
Pseudo/Input=0	-1.97	**-2.12**	-1.39
Pseudo/Input=1	-2.07	**-2.21**	-1.97
Pseudo/Input=2	-2.06	**-2.22**	-1.99
Input/Query=2			
Pseudo/Input=0	**-2.67**	-2.61	-2.40
Pseudo/Input=1	-2.65	**-2.71**	-2.65
Pseudo/Input=2	-2.64	-2.62	**-2.65**
Input/Query=3			
Pseudo/Input=0	-2.72	-2.76	**-3.09**
Pseudo/Input=1	-2.73	-2.76	**-3.21**
Pseudo/Input=2	-2.62	-2.63	**-3.23**

Table 3. Best $log(ANMRR)$ of spatial-scaled pseudo images (positive-and-negative)

	Scheme A: MARS-like	Scheme B: Gaussian-Normalized	Scheme C: Max-val Normalized
Input/Query=1			
Pseudo/Input=0	-1.97	**-2.12**	-1.39
Pseudo/Input=1	-1.95	**-2.17**	-2.03
Pseudo/Input=2	-1.94	**-2.17**	-2.06
Input/Query=2			
Pseudo/Input=0	**-2.67**	-2.61	-2.40
Pseudo/Input=1	-2.60	**-2.71**	-2.67
Pseudo/Input=2	-2.63	**-2.73**	-2.66
Input/Query=3			
Pseudo/Input=0	-2.72	-2.76	**-3.09**
Pseudo/Input=1	-2.89	-2.96	**-3.49**
Pseudo/Input=2	-2.99	-3.05	**-3.51**

Table 4. Best $log(ANMRR)$ of SNR-scaled pseudo images (positive-and-negative)

Negative instances are important, because they tell us about the "undesired" image properties (or image feature values). That is, the user does not want pictures similar to a negative image. However, the negative images do not provide information about a particular feature whether it is good for matching purpose or not. Two negative images can be close or far away, but positive images should always be close together. The simulation results show that negative feedback improves query accuracy in many cases, especially when enough positive instances are given. If the number of input instances is small, only our method can consistently improve the accuracy using the negative instances.

Although both multi-scale schemes that generate pseudo images can enhance the query accuracy (especially for our method), we notice that the SNR multi-scaled images not only produces better performance than the spatial multi-scaled ones, they also have consistently improved results. This may be due to the fact that the spatial-scaled images suffer from the aliasing effect when pictures are down-sampled and thus image features are distorted more than those of the SNR-scaled ones. Overall, Scheme C significantly improves the query accuracy by combining SNR scalability and pseudo-image concepts.

4.3.4. Summary

Based on the above simulation results, we briefly summarize our observations below.

- Distance-based weight estimation outperforms when multiple input instances are available.
- Pseudo images improve query accuracy in many cases, especially when our method is used with SNR scalability.
- Experiments show that one pseudo image per input image gives significant performance boost in many cases.
- Negative instances used as pruning criterion produce better results than those used as negative samples in distance estimation.
- When input instances are few, negative feedback may even degrade the performance of the MARS-like and the Gaussian-normalized schemes.
- When sufficient input instances are available, the Gaussian normalized feature distance does not provide as precise estimation as our method.
- The SNR multi-scaled images provide better ANMRR values; they also lead to more consistent improvements in accuracy.

5. CONCLUSIONS

In this paper, the multi-instance image retrieval problem is investigated. The main contributions of this paper are: (1) propose a distance-based method to estimate user perception based on the given positive instances, (2) generate consistent pseudo images particularly when the query set is too small, and (3) prune irrelevant outcomes based on the given negative images. The first concept is realized by analyzing the scattering level of the query instances in the feature spaces. 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) would have similar numerical values (small scatter numbers) 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. The third one is realized by creating pruning regions in the combined feature space. Our conjecture is that negative instances carry only the information of the undesired image feature values. Namely, undesired images should not look like the negative images because negative images may be close or far away. Thus, they are suitable for pruning rather than for distance estimation.

All the preceding concepts can be integrated into one algorithm using the same basic structure. We examine the performance of our scheme using the ANMRR criterion. 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. As we conjectured, negative feedbacks with pruning scheme performs better than those with weight estimation scheme.

Additional conclusion remarks are (1) the SNR multi-scaled images are generally better than the spatial multi-scaled ones in producing pseudo images; (2) although distance normalization is conceptually useful for weight estimation, simulations show that normalization is not always necessary for producing good results; (3) our method does not require *a priori* information about the data distribution in the database, which not only reduces the computational complexity but also makes it more suitable for searches in a distributed networking environment.

ACKNOWLEDGEMENTS

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.

REFERENCES

1. A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence* **22**, pp. 1349–1380, Dec. 2000.
2. S. Jeong, K. Kim, B. Chun, J. Lee, and Y. J. Bae, "An effective method for combining multiple features of image retrieval," in *IEEE TENCON*, **2**, pp. 982–985, Sep. 1999.
3. 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.* **8**, pp. 644–655, Sep. 1998.
4. Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based image retrieval with relevance feedback in MARS," in *Proc. IEEE Int. Conf. Image Processing*, pp. 815–818, 1997.
5. G. Aggarwal, T. V. Ashwin, and S. Ghosal, "An image retrieval system with automatic query modification," *IEEE Trans. Multimedia* **4**, pp. 201–214, Jun. 2002.
6. F.-C. Chang, H.-M. Hang, and H.-C. Huang, "Research friendly MPEG-7 software testbed," in *Image and Video Communication and Processing Conf.*, pp. 890–901, (Santa Clara, USA), Jan. 2003.
7. T. Ashwin, N. Jain, and S. Ghosal, "Improving image retrieval performance with negative relevance feedback," in *ICASSP*, pp. 1637–1640, May 2001.
8. C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *IEEE Trans. Multimedia* **4**, pp. 260–268, Jun. 2002.
9. *Multimedia Content Description Interface - Part 3: Visual*, *ISO/IEC JTC1/SC29/WG11, FDIS N4203*, MPEG Committee, Jul. 2001.
10. M. Committe, ed., *Subjective Evaluation of the MPEG-7 Retrieval Accuracy Measure (ANMRR)*, *ISO/IEC JTC1/SC29/WG11, M6029*, MPEG Committee, May 2000.
11. B. S. Manjunath, P. Salembier, and T. Sikora, eds., *Introduction to MPEG-7*, John Wiley & Sons Ltd., Baffins Lane, Chichester, West Sussex PO19 1UD, England, 2002.
12. F.-C. Chang and H.-M. Hang, "Content-based image retrieval using both positive and negative feedback," in *International Conference on Multimedia and Expo (ICME)*, (Taipei, Taiwan), June 2004.