Long-Term Cross-Session Relevance Feedback Using Virtual Features

Peng-Yeng Yin, Bir Bhanu, *Fellow*, *IEEE*, Kuang-Cheng Chang, and Anlei Dong

Abstract—Relevance feedback (RF) is an iterative process, which refines the retrievals by utilizing the user's feedback on previously retrieved results. Traditional RF techniques solely use the short-term learning experience and do not exploit the knowledge created during cross sessions with multiple users. In this paper, we propose a novel RF framework, which facilitates the combination of short-term and long-term learning processes by integrating the traditional methods with a new technique called the *virtual feature*. The feedback history with all the users is digested by the system and is represented in a very efficient form as a virtual feature of the images. As such, the dissimilarity measure can dynamically be adapted, depending on the estimate of the semantic relevance derived from the virtual features. In addition, with a dynamic database, the user's subject concepts may transit from one to another. By monitoring the changes in retrieval performance, the proposed system can automatically adapt the concepts according to the new subject concepts. The experiments are conducted on a real image database. The results manifest that the proposed framework outperforms the traditional within-session and log-based long-term RF techniques.

Index Terms—Active nearest neighborhood, content-based image retrieval, long-term learning, relevance feedback, short-term learning, virtual feature.

1 Introduction

Due to increasing demands of managing pictorial data such as art galleries, medical image archiving, trademark signs, etc., the development of efficient image retrieval systems becomes extremely important. Recently, many content-based image retrieval (CBIR) systems have emerged to satisfy some of the needs [1], [2], [3], [4], [5], [6], [7]. Although these modern image databases are queried by image content, the query image still needs to be formulated into some internal form for efficient execution. Since the users, in general, do not know the make-up (what kinds of images are present in the database), the representation (what types of pictorial features (PFs) are used to represent an image), and the techniques (what kinds of indexing methods are employed) used in the search environment, the query formulation process is treated as a series of tentative trials until the target images are found. Relevance feedback (RF) [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] is an interactive process, which fulfills the requirements of the query formulation. Its general idea is described as follows: The user initializes a query session by submitting an image. The system then compares the query image to each image in the database and returns the r images that are the nearest neighbors to the query. If the user is not satisfied with the retrieved result, the user can activate an RF process by identifying which retrieved images are relevant and which are nonrelevant. The system adapts its internal parameter values to include as many user-desired images as possible in the next retrieved result. The process is repeated until the user is satisfied or the results cannot be further improved. The RF techniques provide a way of closing the gap between the *machine subject* in terms of the PFs and the *human subject* who is driven by semantics. The RF paradigm has been shown to be effective in accessing image databases; however, most of the existing RF techniques deal with a single query in a single retrieval session only.

There are two challenging problems when applying the RF techniques to image retrieval. *First*, unlike most applications encountered in computer vision and pattern recognition, it is difficult to deploy supervised learning before the retrieval system is created. The system has no knowledge about which database images are relevant and which are nonrelevant to a set of known labels, since we do not know the user's intention until the user starts the feedback iteration. That is, the system cannot output the retrieval results to a given query based on a sufficiently large set of training data. Instead, the relevant information is collected online via the users' feedback, and this information is very limited, since most users cannot stand too many feedback iterations. *Second*, since image semantics is usually not fully described by the low-level features selected, we need to overcome the disagreements between human subjects and machine subjects.

In this paper, we propose a novel RF approach for CBIR. In addition to taking the feedback information collected within a single query session into account, we further digest the long-term feedback history collected from multiple query sessions and represent the derived metaknowledge in a very efficient form called the *virtual feature* (VF). Based on the VF, we can estimate the semantic similarity between the query image and a database image. This estimate is then used to define an *active* (dynamically adjusted) nearest neighborhood for the query image. When the relevant

- *P.-Y. Yin and K.-C. Chang are with the Department of Information Management, National Chi Nan University, 470 University Rd., Puli, Nantou 54561, Taiwan. E-mail: {pyyin, s3213518}@ncnu.edu.tw.*
- *B. Bhanu and A. Dong are with the Center for Research in Intelligent Systems, Bourns College of Engineering, University of California at Riverside, 216 EBU II, BCOE, Riverside, CA 92521. E-mail: bhanu@cris.ucr.edu, anlei.dong@gmail.com.*

Manuscript received 25 Oct. 2006; revised 12 Aug. 2007; accepted 24 Sept. 2007; published online 3 Oct. 2007.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0497-1006.
Digital Object Identifier no. 10.1109/TKDE.2007.190697.

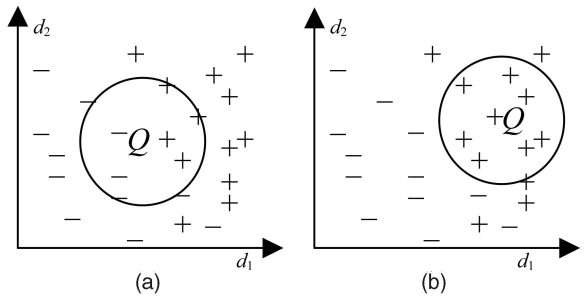


Fig. 1. QVM. (a) The original query. (b) The query is moved to a region that involves more relevant images. "+": relevant, "−": nonrelevant, and "Q": query.

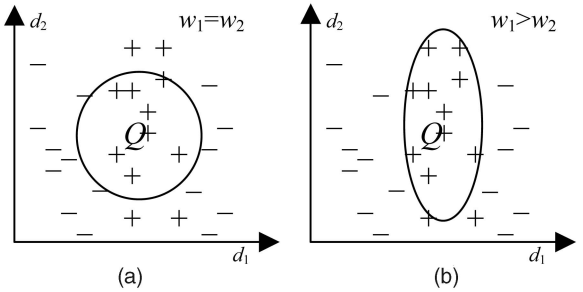


Fig. 2. FRE. The nearest neighborhood boundary forms (a) a circle with equal relevances and (b) an ellipse with different relevances. "+": relevant, "−": nonrelevant, and "Q": query.

concept associated with a particular query is changed, the proposed system can adapt to learn the new concept. The proposed method provides an effective and efficient framework, which integrates the traditional short-term feedback learning and the newly introduced long-term feedback learning.

The remainder of this paper is organized as follows: Section 2 reviews the related research on RF and long-term learning and identifies the contributions of this paper. Section 3 describes the proposed method. Section 4 presents the experimental results and provides comparative performance evaluation. Finally, Section 5 presents the conclusions of this paper.

2 Related Work and Contributions of This Paper

Let the query image and a database image be represented by the PF vectors $Q = (q_1, q_2, \ldots, q_t)$ and $D = (d_1, d_2, \ldots, d_t)$, respectively, where t is the number of selected features, and q_i and d_i are the representative values of the ith feature. The CBIR systems derive the similarity by computing a distance function between Q and D. The normalized euclidean metric,

$$Dist(Q, D) = \sqrt{\sum_{i=1}^{t} \Delta_i^2 \Big/ t},$$

where $\Delta_i = q_i - d_i$, is generally used for this purpose. Qamra et al. [18] proposed the dynamic partial function (DPF)

$$Dst(Q, D) = \sqrt{\sum_{\Delta_i \in \Omega} \Delta_i^2 \Big/ t},$$

where Ω is the set of the m smallest values of Δ_i $(i = 1, 2, \ldots, t)$ to activate different features for different image pairs. They empirically showed that DPF models human perception better than other distance functions do. The major RF techniques for image retrieval include *query vector modification* (QVM) [8], [9], *feature relevance estimation* (FRE) [10], [11], [12], and *classification-based* (CB) *methods* [13], [14], [15], [16], [17]. The QVM method reformulates the query vector through user's feedback by

$$Q = \alpha Q + \beta \sum_{D_j \in R} \frac{D_j}{|R|} - \gamma \sum_{D_j \in N} \frac{D_j}{|N|},$$

where D_j are images that belong to relevant set R or nonrelevant set N, and α, β, and γ are the weights so as to move the query toward relevant images and away from nonrelevant ones (see Fig. 1). The FRE approach learns the weight w_i for each low-level feature d_i and computes the dissimilarity by

$$Dist(Q, D) = \sqrt{\sum_{i=1}^{t} w_i (q_i - d_i)^2}.$$

Please see Fig. 2. The CB method trains a classifier (based on SVM, boosting, or Bayesian classifier) from the prior history of feedbacks for classifying the test data. Each category of methods has its own strengths and weaknesses. The disadvantages of these techniques are summarized in Table 1.

Most of the approaches refine the retrieved results based on the feedback history within one query session only. A *query session* includes the period that consists of the submission of the original query and all the subsequent feedback iterations. Hence, these approaches maintain a form of short-term memory that captures the user's intention for this specific query session, and this knowledge is dropped before the next query session is initiated. There is no consideration given for the *cross-session* feedback history. For a specific-domain image database, for example, the trademark database, users are domain experts whose relevant concepts are highly consistent in order to justify whether the submitted images cause infringement to a registered trademark. For a general image database, the consistency ratio could be lower, but the long-term experience that captures the common agreement among various query sessions fulfilled by the same or different users is useful for future queries in the convergence of the feedback process to occur at a faster rate. Several long-term learning methods have been proposed [19], [20], [21], [22], [23], [24], [25]. The proposed VF approach and the existing long-term learning methods can be classified as shown in Table 2.

2.1 Contributions of This Paper

In comparison with the state-of-the-art approaches for CBIR shown in Table 2, the fundamental contributions of this paper are:

TABLE 1
Disadvantages of Major RF Techniques

Approaches	Disadvantages
QVM	• Not every relevant image is consistently relevant to the query along every feature dimension. • Assumes that the distribution of the locations of the relevant images forms an intrinsic cluster and the clustering result is valid for the chosen distance function. • The feature vectors of the relevant images are averaged to compute the new query. Thus, some information is lost regarding the identity of relevant images.
FRE	• The query cannot be moved toward a more desired region of the feature space. • Relevant images may not be selected in the neighborhood of a query. • The identity of relevant images is not stored, only the feature relevance is computed.
CB	• Needs more training examples to well train a classifier. • On-line training of a classifier could be time-expensive.

TABLE 2
Classification of the Proposed VF Approach and the Existing Long-Term Learning Methods

Approaches \ Features	Image-based feedback	Region-based feedback	Combine user feedback with low-level features	User feedback and low-level features are used separately	User feedback is used for image grouping	User feedback is used for model selection	Gradually discount the previous feedback
VF (this paper)	•		•				•
Minka & Picard [19]		•		•	•	•	
Bhanu & Dong [20]	•			•	•		
He et al. [21]	•			•			
Jing et al. [22]		•	•				•
Dong & Bhanu [23]	•			•		•	
Yin et al. [24]	•		•			•	•
Hoi & Lyu [25]	•		•				

1. A novel technique that integrates the short-term and the long-term learning experiences. The short-term experience is captured using the existing RF methods and the long-term experience is learned through a new mechanism called the *VF*. The long-term relevant information is stored in the VF of each image by digesting the RF history. This information is then used to dynamically adjust the neighborhood boundary of the query. The method inherits the advantages of existing RF approaches and can attain a higher average precision.
2. A mechanism that updates the concepts associated with a particular image by adjusting the weights between the latest and the aged information.
3. An adaptive dissimilarity measure based on VF and its convergence analysis for concept learning.
4. Detailed experimental results that deal with all aspects of the approach and a comparison with [25] on a large real-world image database.

These experiments convincingly demonstrate the contributions and the performance gains achieved by the proposed approach.

3 The Proposed Approach

To digest the relevant information accumulated from within-session or cross-session query experiences, we add a VF to the feature vector of each of the database images. The value of the VF is determined by the set of relevant images specified by various users. The VF is then used to assist the original PFs to evaluate the degree of similarity between the query and database images in accordance with the human subject.

3.1 Notation and Terminology

Assume that the retrieval system is designed to process many query sessions from multiple users. In each query session, the user can activate an RF process for a number of iterations to improve retrievals. We denote by Q_i the query of the ith query session and by R_j the set of relevant images identified at the jth feedback iteration during the current query session. To improve the retrievals, the feature vector of query Q_i is repeatedly reformulated from the feature vectors of the relevant images in R_j. The feature vector of a database image D consists of two parts, the PFs and the VF, denoted by $PF(D)$ and $VF(D)$, respectively. The $PF(D)$ is

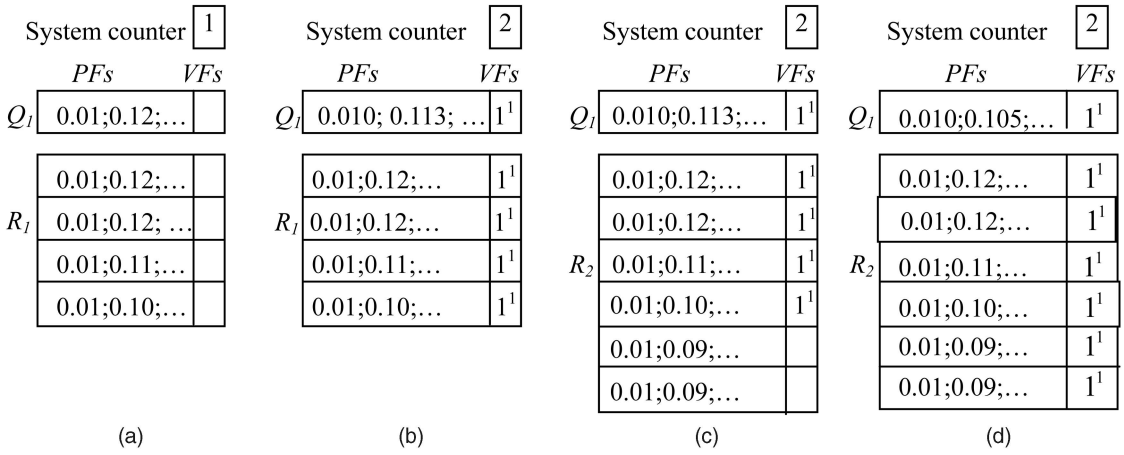


Fig. 3. Example 1: An illustrative example of the VF updating at the first query session.

a vector of numerical values that are computed via image processing and computer vision techniques, whereas the $VF(D)$ is a sequence of symbols that are obtained using the VF updating algorithm, as noted in the next section.

The user implicitly identifies a concept by marking a set of relevant images in his/her feedback. Because different concepts may be sought at different query sessions, an image can possess multiple concepts. The VF of an image records all of the relevant concepts and their significance related to this image. It is given as $VF(D) = c_1^{e_1} \otimes c_2^{e_2} \otimes \cdots \otimes c_m^{e_m}$, where c_i is the index value of the concept, e_i is the support value for the corresponding concept, and $\otimes$ is a delimiter for separating the terms for individual concept. The value of c_i is a positive integer, and the support value e_i is a nonnegative real value. In addition, we define $\left(c_1^{e_1} \otimes c_2^{e_2} \otimes \cdots \otimes c_m^{e_m}\right)^k = c_1^{ke_1} \otimes c_2^{ke_2} \otimes \cdots \otimes c_m^{ke_m}$ for any $k \geq 0$ and $c_i^0 \equiv Null$ (that is, c_i is removed from the corresponding VF). The vector that consists of support values $(e_1, e_2, \ldots, e_m)$, referred as e-vector, indicates the relative frequencies that image D is identified as relevant to those concepts. A transformation function is used to transform a VF into a canonical form $VF(D) = c_1^{e_1} \otimes c_2^{e_2} \otimes \cdots \otimes c_m^{e_m}$, where $c_i < c_j$ if $i < j$ and the e-vector satisfies the constraint $\sum_{i=1}^{m} e_i^2 = 1$. The transformation function *transform(VF(D))* consists of three steps:

- Step 1 (Reordering). Reorder the sequence of $VF(D)$ in nondecreasing order of the concept index. For example, the VF $5^{0.12} \otimes 8^{0.26} \otimes 5^{0.17} \otimes 3^{0.66} \otimes 8^{0.11}$ is reordered as $3^{0.66} \otimes 5^{0.12} \otimes 5^{0.17} \otimes 8^{0.26} \otimes 8^{0.11}$. The reordering process provides the convenience for accumulating the support values of the same relevant concept as noted in the next step.
- Step 2 (Merging). Merge multiple terms of the same concept by summing their support values. For example, the VF $3^{0.66} \otimes 5^{0.12} \otimes 5^{0.17} \otimes 8^{0.26} \otimes 8^{0.11}$ is modified as $3^{0.66} \otimes 5^{0.29} \otimes 8^{0.37}$ after the merging process. Thus, the support values related to the same concept are aggregated.
- Step 3 (Normalization). Finally, the e-vector of the VF is normalized to a unit vector by constraining $\sum_{i=1}^{m} e_i^2 = 1$. For example, the VF $3^{0.66} \otimes 5^{0.29} \otimes 8^{0.37}$ is changed to $3^{0.81} \otimes 5^{0.36} \otimes 8^{0.46}$ after the normalization process. It is noteworthy that every image is treated as a whole, and the support value for various concepts is defined as a unit vector on the unit sphere. Normalizing the e-vector will not affect the ratio between support values for different concepts but will simplify the computation of our similarity measure (which is based on the cosine distance) by performing inner product of e-vectors.

The transformation function guarantees that the VF that results from the operations performed on multiple VFs is still in the canonical form.

3.2 Virtual-Feature Updating Algorithm

Unlike traditional PFs, which represent the low-level visual features of the images, the proposed VFs capture the high-level semantics imposed by multiple users. The semantic concepts are learned from prior retrieval experiences *in breadth* (within-one-session learning with a single query) and *in depth* (within-multiple-session learning with multiple queries). For the in-breadth learning, the user can guide the system to increase the weights (support values) of the concepts being sought within a particular query session by requesting as many feedback iterations as he/she wishes. For the in-depth learning, the system automatically derives the long-term knowledge based on previous retrieval experiences with multiple users, and this knowledge is then used to expedite the search for relevant images in future query sessions. The proposed VF technique, facilitating both types (in breadth and in depth) of learning, demonstrates high efficacy in real-world applications.

For the convenience of presentation, we first show two illustrative examples and then provide the VF updating algorithm at the end of this section. These examples demonstrate how the VFs for query and a database image are computed and updated.

Example 1. The first example, as shown in Fig. 3, illustrates the beginning scenario of the VF learning process. Initially, every database image is represented by a feature vector that consists of PFs only. We add an empty VF to the feature vector of each image. When the first query session is activated by a user and letting $Q_1 = (0.01; 0.12, \ldots)$, the retrieval system searches the top-r-nearest images by using the distance function defined by the adopted short-term RF technique, for example, the QVM method. If the user is not satisfied

with the result, he/she would activate an RF process. Let R_1 be the set of identified relevant images at the first feedback iteration (see Fig. 3a). Since all relevant images in R_1 carry empty VFs (no a priori relevant information exists at this moment), they derive the VFs by requesting a number from the system counter. The value of the system counter serves as the index of a particular relevant concept. It starts counting from 1 (indicating the index of the first relevant concept experienced by our system), and it is incremented by 1 every time a request is made (that is, indicating the index of a new relevant concept identified in the most recent feedback marked by the user). It is used to obtain the identity of the relevant concept learned from the RF history. As such, the semantic similarity between images can be derived. Therefore, as Fig. 3b shows, all the images in R_1 are assigned the value 1^1 as their VFs to mark that these images possess the first relevant concept experienced by the system.

To utilize the relevant information, the query Q_1 is reformulated based on R_1 for the next retrieval iteration. The PF part of Q_1 can be updated as the average PFs of images from R_1, that is, $PF(Q_1) = (0.010; 0.113, \ldots)$, according to the QVM method. Analogously, we let Q_1 derive its VF by concatenating the VFs of all images from R_1, viz.,

$$VF(Q_1) = transform\Big(VF(D_1) \otimes VF(D_2) \otimes \cdots \otimes VF\Big(D_{\big|_{R_1}\big|}\Big)\Big),\ D_k \in R_1. \tag{1}$$

In particular, for the example in Fig. 3b, the VF of Q_1 first gets its intermediate form as $\left(1^1 \otimes 1^1 \otimes 1^1 \otimes 1^1\right) = 1^4$, then transforms the support value to a unit vector, and gets the final form as 1^1. The feature vector of Q_1 thus becomes $(0.010; 0.113; \ldots; 1^1)$ and is used to retrieve the new top-r-nearest images (see Section 3.3). Let R_2 be the relevant set at the second feedback iteration, where another two relevant images with empty VFs are introduced, as shown in Fig. 3c. Let $R_2' = \{I | I \in R_2 \text{ and } VF(I) = \text{Null}\}$ and $R_2'' = \{I | I \in R_2 \text{ and } VF(I) \neq \text{Null}\}$. Each of the images in R_2' first derives the preliminary form of VF by learning the a priori relevant information from the images in R_2'' to mark that the relevant images in R_2' should also own the concepts already learned in the long term by the relevant images in R_2''. This can be done by

$$VF(I) = transform\Big(VF(D_1) \otimes VF(D_2) \otimes \cdots \otimes VF\Big(D_{\big|_{R_2''}\big|}\Big)\Big),\ D_k \in R_2'', \text{for any } I \in R_2'. \tag{2}$$

In practice, we confine the maximal length of the VF for a database image to be 200, which is a reasonable estimate for the maximal number of concepts contained in an image. When the VF of an image is reformulated, we remove the VF concepts with the least significant support values if the VF length exceeds 200. Here, in our example, the two relevant images at the bottom of Fig. 3c will derive the preliminary form of VF as $\left(1^1 \otimes 1^1 \otimes 1^1 \otimes 1^1\right) = 1^4$ and transform it to 1^1. At the current display, if all relevant images in R_2 have already shared one common VF concept, then we can directly perform the query reformulation (1), as shown in Fig. 3d. Otherwise, for every relevant image in R_2, the corresponding VF is further updated by incorporating the a priori relevant information with the *current* relevant information in a weighted product fashion:

$$VF(I) = transform\Big(\big(VF(I)\big)^{\eta} \otimes counter^{(1-\eta)}\Big),\ 0 \leq \eta \leq 1,\ \forall I \in R_2, \tag{3}$$

where *counter* is the current value of the system counter and denotes the index of the new relevant concept, and η is the weight that controls the importance of each component, and it is adaptively tuned, as noted later in this section. Thus, all relevant images at the current iteration share a common concept by requesting a number from the system counter to mark that all these images are simultaneously identified as relevant at the current iteration (the latest relevant information), and this new concept is incorporated into previous VFs (the aging relevant information) in a weighted manner. With this mechanism, the VF technique can facilitate adaptive concept learning that is essential in the dynamic environment due to the deletions and insertions of database images. Furthermore, we keep only one VF for an image that is learned when the image is identified as a relevant image. Note that the query image is treated as a duplicate of the database image, and the VF of the query image is temporarily used in the current session only for capturing the user's intention, and it is not stored for future sessions. The VF of the query image should not be associated with the VF of the original database images, since it maintains the consensus concepts.

Example 2. Fig. 4 shows a more complex example of another query session, say, the 12th query session, and let $Q_{12} = (0.02; 0.11; 0.05, \ldots)$ with system counter value currently be equal to 34 (that is, the system has generated 33 relevant concepts based on the previous interactions with the users). The relevant set R_1 identified at the first feedback iteration may already contain some images with multiple concepts learned from previous query sessions (long-term history), and some images may not have VFs yet, since they had never been identified as relevant at any previous sessions (see Fig. 4a). These latter images first learn their VFs from the a priori relevant information carried by other relevant images in R_1 by using (2) and obtain the preliminary VFs form as $1^{0.69} \otimes 2^{0.17} \otimes 13^{0.69} \otimes 14^{0.17}$. Then, the VFs of all images in R_1 are further updated using (3) to incorporate the latest relevant information represented by concept 34, and assuming $\eta = 0.8$, the updated VFs are shown in Fig. 4b. The query vector is reformulated as $Q_{12} = (0.014; 0.106; 0.052, \ldots, 1^{0.67} \otimes 2^{0.17} \otimes 13^{0.67} \otimes 14^{0.17} \otimes 34^{0.28})$ to continue the subsequent search.

3.2.1 Adaptive Tuning of η

As mentioned previously, η controls the relative importance between the aging and the latest relevant information. It can dynamically be adapted to facilitate the relevant concept transition for a particular image. Considering a dynamic database, where the deletions and

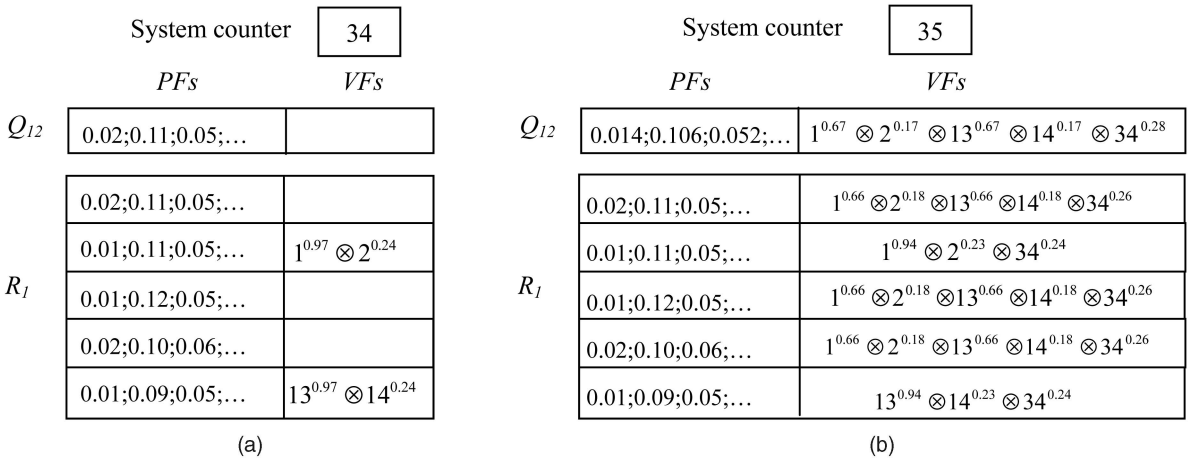


Fig. 4. Example 2: An illustrative example of the VF updating at the 12th query session.

insertions of database items occasionally happen, a user's subjective judgment with regard to some of the images (that are related to the items deleted or inserted) would also change, since he/she expects to see new relevant images to be retrieved when those images are used as queries. Thus, the transition between concepts is query dependent. We will know whether there is an ongoing concept transition for a particular query image by monitoring the changes in its retrieval performance. We measure the retrieval performance for query Q_i by using the precision rate (PR) defined as $Precision(Q_i) = \frac{\text{Number of positive retrievals}}{\text{Number of total retrievals}} \times 100\%$ via the user's feedback on image retrievals. If the current precision $Precision_{cur}(Q_i)$ is greater than the previous precision $Precision_{pre}(Q_i)$ (the one obtained at the immediate preceding iteration or the last precision obtained when Q_i was previously used as a query), then the current concept is in consensus with the a priori relevant information (carried by previous VFs), and the importance of previous concepts should be enhanced by increasing the value of η. On the other hand, if $Precision_{cur}(Q_i) < Precision_{pre}(Q_i)$, then it means that the current concept is contrary to the previous one, there is a concept transition happening, the contribution of previous VFs should be discounted (by decreasing the value of η), and the importance of adding a new concept is increased. With these considerations, we adaptively tune $\eta(Q_i)$ for a particular query according to

$$\eta(Q_i) = \eta(Q_i)(1 + Precision_{cur}(Q_i) - Precision_{pre}(Q_i)), \tag{4}$$

and the tuned value should be bounded by $0 \leq \eta(Q_i) \leq 1$.

Note that the PR mentioned above is the traditional definition: it is only concerned with the number of relevant images that appear in a fixed number of retrievals, and it does not individually account for the rank of relevant images. In the experimental result section, we use a new definition of average precision that individually accounts for the rank of all relevant images in the database, and as such, the capability of our system is more accurately evaluated.

VF updating algorithm. The VF updating algorithm is shown in Fig. 5. In summary, the VF technique derives multiple semantic concepts by three means. *First*, the newly seen relevant image with empty VF learns the preliminary form by digesting the VFs from other relevant images on a normalized basis. This is an efficient way for a newly seen relevant image to learn the a priori consensus relevant information reached in previous query sessions. The retrieval system can thus attain a very high performance, even at early feedback iterations if the user is seeking for relevant concepts that are similar to those sought by the majority of previous users. *Second*, all relevant images identified at the current iteration learn a new concept that is sought at this particular iteration, and this concept is incorporated into previously learned ones in a weighted manner. *Third*, the relative importance between the a priori relevant information and the latest relevant information is adaptively tuned by monitoring the transitions of concepts. Thus, if the retrieval performance does not reach a satisfactory level at earlier query sessions, the VF of relevant images (and also of the query) will be prone to emphasize the new concepts that are learned at the latest iterations to accomplish the relevance transition. The VF for a query image is used during the in-breadth session only, and it is not stored for future use.

3.3 Dissimilarity Measure

The proposed dissimilarity measure that is used in the computation of r-nearest images will be presented in this section. Let the VF of an image D be $c_1^{e_1} \otimes c_2^{e_2} \otimes \cdots \otimes c_m^{e_m}$. We first define the concept set of image D as $C = \{c_1, c_2, \ldots, c_m\}$, each concept c_i being associated with a support value e_i. The larger the cardinality of the concept set, the more general the overall concept delivered by the image. In addition, the larger the support value, the more important the corresponding concept. In this paper, the semantic similarity between a query Q and an image D is defined as the cosine value of the angle θ between the e-vectors of Q and D. Since the e-vector is defined on the unit sphere ($\sum_{i=1}^{m} e_i^2 = 1$), the cosine value can be obtained by simply computing the inner product of e-vectors. The semantic similarity is used to adjust the distance $Dist(Q, D)$ calculated by PFs. We finally define a dissimilarity measure $Dist_{VF}(Q, D)$ based on both semantic and pictorial information as follows:

System initialization

1. Add an empty VF to the feature vector of each database image.
2. Set *counter* = 1 and $i = 0$.

Query session

3. While a query is submitted by a user Do
   - 3.1. $i = i + 1$.
   - 3.2. Let Q_i be current query, set j =1.
   - 3.3. Compute r nearest images using the proposed dissimilarity measure (Eq. (5)). The tie is broken according to the distance in the PF space.
   - 3.4. While user is not satisfied with the retrieved result Do
     - (a) User marks the r images as relevant or nonrelevant.
     - (b) Denote by R_j the set of relevant images, and by N_j the set of nonrelevant images. $R_j' = \{I \mid I \in R_j \text{ and } VF(I) = \text{Null}\}$, and $R_j'' = \{I \mid I \in R_j \text{ and } VF(I) \neq \text{Null}\}$.
     - (c) If $R_j'' = \varnothing$ then $VF(I) = counter$ for each $I \in R_j$ and go to step (g)
     - (d) For each image $I \in R_j'$ Do
       $VF(I) = transform\,(VF(D_1) \otimes VF(D_2) \otimes \cdots \otimes VF(D_{|R_j''|})), \quad D_i \in R_j''$.
     - (e) $\eta(Q_i) = \eta(Q_i)\,(1 + Precision_{cur}(Q_i) - Precision_{pre}(Q_i))$ bounded by $0 \le \eta(Q_i) \le 1$
     - (f) For each image $I \in R_j$ Do
       $VF(I) = transform\left((VF(I))^{\eta(Q_i)} \otimes counter^{(1-\eta(Q_i))}\right)$
     - (g) Update the PF of the query
       $Q_i = \alpha Q_i + \beta \sum_{D_k \in R_j} D_k / |R_j| - \gamma \sum_{D_k \in N_j} D_k / |N_j|$
     - (h) Update the VF of the query
       $VF(Q_i) = transform\ (VF(D_1) \otimes VF(D_2) \otimes \cdots \otimes VF(D_{|R_j|})), \quad D_i \in R_j$
     - (i) set $j = j + 1$ and *counter* = *counter* +1
     - (j) Compute r nearest images using the proposed dissimilarity measure (Eq. (5)).
   - end

end

Fig. 5. The VF updating algorithm.

$$Dist_{VF}(Q,D) = \begin{cases} (1-\cos\theta)Dist(Q,D), & \text{if both } VF(Q) \text{ and } VF(D) \\ & \text{are known,} \\ Dist(Q,D), & \text{otherwise,} \end{cases} \quad (5)$$

where $0 \le \cos\theta \le 1$, since e-vectors are defined in nonnegative real vector space. Note that θ could be zero for multiple images. As a result, $Dist_{VF}(Q,D)$ will be zero for multiple images. In these situations, we break this tie with $Dist(Q,D)$, which is the distance in the PF space. With the definition in (5), the distance $Dist_{VF}(Q,D)$ is reduced if the e-vectors of Q and D are not orthogonal to each other $(\cos\theta > 0)$, and it is not changed if the e-vectors of Q and D are orthogonal or at least Q or D has not yet derived a VF. Note that the distance reduction is a relative quantity, depending on how semantically similar the e-vectors of Q and D are. The distance can even be reduced to zero when Q and D have identical e-vectors for the corresponding concepts, so two dissimilar images in the PF space but are semantically alike can be retrieved together by using our VF mechanism, as will be seen in the retrieval examples in Section 4. Therefore, the proposed method dynamically adjusts the distance between the query and the database images based on the VFs, which are derived from the long-term feedback history.

Our method provides a framework for integrating the short-term and the long-term RF techniques in a single retrieval system. The short-term RF technique is implemented using the PF part of the query, and the long-term RF digestion is carried out by the VF part of the query. For example, if the QVM is adopted for the short-term RF, the advantages of the proposed method can be illustrated in Fig. 6. *First*, the query vector can be moved by the QVM toward the subspace that contains more relevant images. *Second*, an active nearest neighborhood of the query, which is dynamically defined by the VFs, can lead to a more desired result.

3.4 Concept Learning and Convergence Analysis

The VF technique can be used for concept learning. Let C_1 and C_2 be the two concept sets of images. We define the generalized concept set of C_1 and C_2 as $C = C_1 \cup C_2$ if $C_1 \cap C_2 \neq \emptyset$. That is, if two images have at least one concept in common, their generalized concept set is the union of the respective concept sets. We can augment the size of the generalized concept set by examining the concept set of every image in the database and finally obtain a maximally generalized concept set (MGCS). There may be several MGCSs $C_{\max_1}, C_{\max_2}, \ldots, C_{\max_r}$ that exist in a database such that $C_{\max_i} \cap C_{\max_j} = \emptyset$ for $i \neq j$. If we assume that the

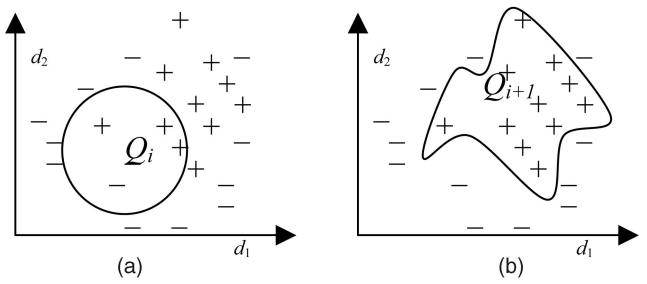


Fig. 6. Illustration of the method that integrates the QVM and the VF techniques. (a) The original nearest neighborhood boundary forms a circle. (b) The query vector is moved by the QVM toward the subspace that contains more relevant images, and the active nearest neighborhood derived by the VFs can lead to a more desired result.

database can be partitioned into a number of disjoint clusters according to human preferences, every MGCS would represent a particular image category (a finer partition of the true clusters), where each member of the category possesses at least one concept, which is also contained in the corresponding MGCS.

There are a number of potential applications of the MGCS:

1. *Category browsing.* By deriving all the MGCSs in a database, a directory of image categories could be provided for browsing. The user can quickly obtain the knowledge of the database make-up by using the category browsing before he/she makes a specific query.
2. *Retrieval speedup.* If the query image is selected from the same database, considerable savings of the computations can be achieved. This is done by comparing the concept set of the query image with all MGCSs of the database and determining which image category the query should be matched.
3. *Concept learning.* If we assume that the database can be partitioned into a number of disjoint clusters according to human preferences, the MGCSs can provide an approximation to the clustering result. The accuracy of the approximation is dependent on the allowed number of retrieved images that are presented to the user at every iteration and on the number of experienced query sessions (NQ).

In the following, we show that concept learning can be used to experimentally verify the convergence speed of the VF approach. Assuming that there are c semantic classes that can well partition a database of S images, there must be a ground-truth binary partition matrix $G = \{g_{ij}\}$, $1 \le i \le S$, $1 \le j \le c$, and $g_{ij} = 1$ means that the ith database image belongs to the jth class. Otherwise, $g_{ij} = 0$, subject to the constraints $\sum_{j=1}^{c} g_{ij} = 1$ (each image belongs to exactly one class) and $\sum_{i=1}^{S} g_{ij} \neq 0$ (each class contains at least one image). Likewise, we can compute a partition matrix based on the MGCSs derived from the VFs. Let v be the NQ. The partition matrix is denoted by $M^{(v)} = \{m_{ij}^{v}\}$, $1 \le i \le S$, $1 \le j \le c'$, where c' is the number of MGCSs derived at the vth query session. We can analyze the convergence property of the VF approach by verifying whether $M^{(v)}$ converges to the ground-truth matrix G as v increases. Since $M^{(v)}$ and G are sparse matrices, and they, in general, have different numbers of entries, we do not intend to compute any norm between the two matrices. Instead, we assess the agreement on the class assignments for every pair of distinct images. Considering two different database images, there are four possible cases on the clustering agreement:

1. Both G and $M^{(v)}$ determine that the two images are in the same cluster (S_{11}).
2. The ground-truth G assigns the two images in different clusters, but $M^{(v)}$ positions them in the same cluster (S_{01}).
3. The ground-truth G assigns the two images in the same cluster, but $M^{(v)}$ puts them in different clusters (S_{10}).
4. Both G and $M^{(v)}$ determine that the two images are in different clusters (S_{00}).

We calculate the frequencies of the four possibilities by considering all pairs of different images from the database; that is, if the database size is S, the summation of $S_{11} + S_{01} + S_{10} + S_{00}$ is equal to $S(S-1)/2$. The frequencies of S_{11} and S_{00} reveal the agreement for image relevance and image nonrelevance, respectively, whereas the frequencies of S_{01} and S_{10} depict the disagreement. Because the frequency of S_{00} is significantly greater than that of S_{11}, the nonrelevance agreement will dominate the overall agreement ratio. We measure the relevance agreement (RA) between the two subjects G and $M^{(v)}$ by

$$RA(G, M^{(v)}) = \frac{S_{11}}{S_{11} + S_{10} + S_{01}} \times 100\%. \tag{6}$$

Thus, we can determine the convergence speed of the VF approach by verifying how fast $RA(G, M^{(v)})$ grows as v increases. More details about VF learning rate (VFLR) are given in Section 4.3.

4 Experimental Results

We have implemented the QVM [8], FRE [11], SVM [15], and the proposed VF technique for comparative performance evaluation. We adopted the *experimental design* technique to select the optimal values of the QVM parameters α, β, and γ. The feasible value of each parameter is set to $0, 0.1, 0.2, \ldots, 1.0$, respectively. All combinations of the feasible values are tested subject to the constraint $\alpha + \beta + \gamma = 1$. In our case, it required 66 experiments, and the parameter values with the best performance among these experiments are $\alpha = 0.1$, $\beta = 0.8$, and $\gamma = 0.1$. The parameter η (see (3)) involved in the VF updating algorithm is adaptively tuned automatically.

4.1 Real Image Database

The image relevance is automatically determined by checking whether the returned images belong to the same (manually defined) class as the query. The ground truth is used only to evaluate the performance: it is *not* assumed to preexist in using our VF approach. Thus, the proposed method can deal with a "never-seen-before" database. The

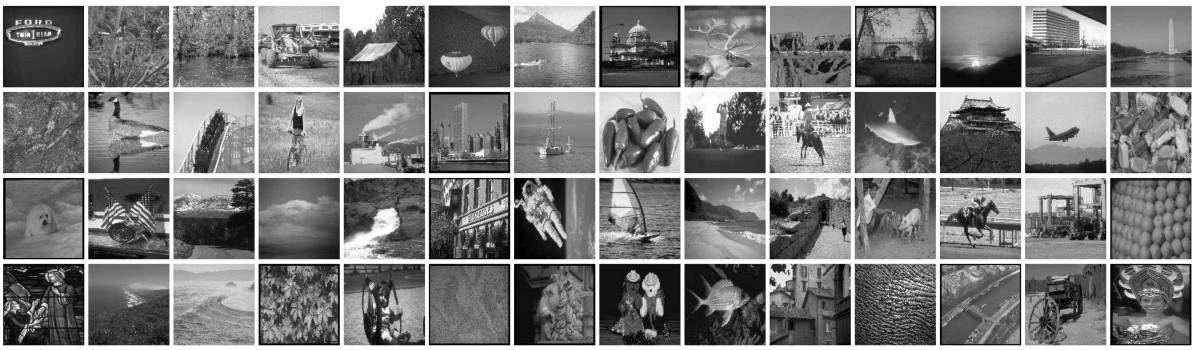


Fig. 7. Sample images from 56 classes (arranged from left to right and from top to bottom) of the UCR database.

UCR database is obtained from the UCR Web site [26]. There are 10,038 images that cover a variety of real-world scenes such as castles, cars, humans, animals, etc. In order to conduct an automatic evaluation of extensive experiments, the database images are manually prelabeled, and the annotations are used to determine the relevance relationships. Three persons are asked to make independent judgment regarding the class assignment, with 56 possible class names given to them. Each image is assigned to the appropriate class when the majority of these three persons agreed with this assignment. For the images that are assigned to three distinct classes by the three persons, the labeling provided by the image processing and computer vision expert (having more than 2 years of experience in the field) is used as the ground truth. The sample images from each class are shown in Fig. 7. The number of images in each class varies from 20 to 695. Each image in the database is represented by a 22-dimensional feature vector that has 16 Gabor features [11] (the mean and standard deviation of filtered images at four orientations and two scales) and six color features (the mean and standard deviation from the HSV color domain [27]).

4.2 Comparative Performance Evaluation

To simulate the practical situation of online users, the sequence of query images used in all the experiments is generated at random. At each query session, the system refines its retrievals by executing the embedded short-term RF technique for several iterations. We use the *Average Precision* metric from the NIST TREC video (TRECVID) as our performance measure. It is defined as the average of precision values obtained after each relevant image is retrieved. Let $\overline{P}$ be the average precision, and it is computed by $\overline{P} = \sum_{D_i \in R} P_i / |R|$, where P_i denotes the precision value obtained after the system retrieves the top-ranked i images, D_i is one of the relevant images, R is the set of all relevant images in the database, and $|R|$ is the cardinality of R. As an example, assume that the database images belong to a number of classes. One of the classes consists of four relevant images, and our retrieval system ranks these relevant images first, second, fourth, and seventh. Thus, the precision values obtained when each relevant image is retrieved are 1, 1, 0.75, and 0.57, respectively. The average precision computes the average of these precision values, and it is 0.83. The average precision over all relevant images can avoid precision fluctuation that is usually encountered by the traditional precision measure. The average PR is calculated at all different displays, namely, the one without any RF interaction (PR0), the one after the first iteration of RF (PR1), the one after the second iteration of RF (PR2), and so forth. We calculate the *average PR* at a query session that is achievable with any query image in the database. For example, at the $5,000\text{th}$ query session, we calculate the average PR by using every database image as a query. The query session is used as a time stamp. The average PR is the performance measure that evaluates the capability achieved by our system at that time stamp. Thus, the VF learning is not performed during this evaluation to make our system stay at a constant performance level within this period. The comparative performances are analyzed for both in depth (with the NQ) and in breadth (with the number of feedback iterations (NF)).

4.2.1 Retrieval Average Precisions with Query Sessions Increased

For the in-depth performance analysis, the average precision with various NQs, using the QVM *without* the VF approach (10 retrieved images are presented at each iteration), is carried out. The future retrieval performance does not improve as NQ increases, since the QVM conducts short-term learning, and the metaknowledge derived from current query session is not used in subsequent query sessions. The average PRs at different displays are $\text{PR0} = 24.7\%$, $\text{PR1} = 39.2\%$, and $\text{PR2} = 40.5\%$. Similarly, for the FRE *without* the VF approach, the average PRs at different displays are $\text{PR0} = 24.7\%$, $\text{PR1} = 39.5\%$, and $\text{PR2} = 41.4\%$. Fig. 8a shows the results when QVM is combined with the VF method. The average PRs rapidly climb up as NQ increases due to the contribution of the use of the active nearest neighborhood facilitated by the VF technique. The users' preferences are well described by the learned concepts stored in the VFs. The converged average PRs at PR0, PR1, and PR2 are 71.4 percent, 77.5 percent, and 78.3 percent, respectively, which are much higher than the ones obtained by the QVM without the VF method. This is because the relevant information learned from the previously queries provides a valuable clue for reformulating the response to future instances of these queries. The VF approach is a general framework that lets the short-term RF technique benefit from long-term learning. Fig. 8b is another example that illustrates the average precisions obtained when SVM [15] is combined with the VF method. It is observed that the long-term learning conducted by the VF approach can also help SVM in improving the average

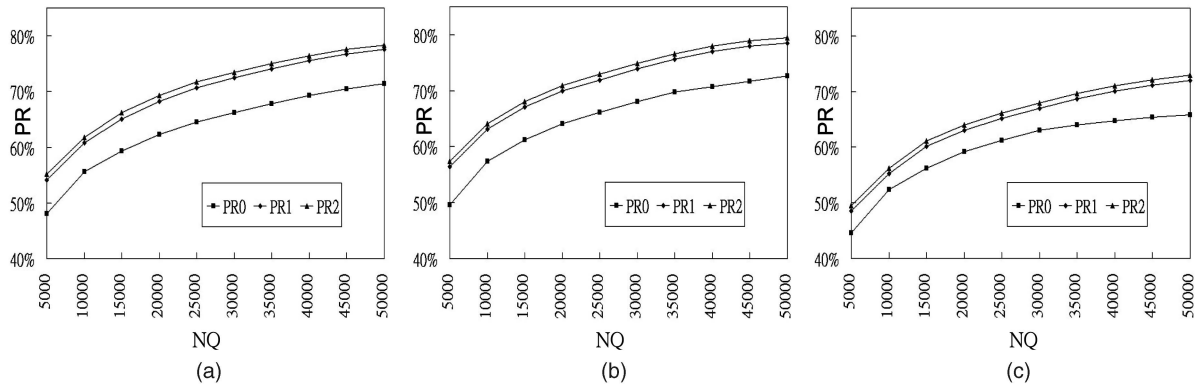


Fig. 8. Average PRs versus NQ. (a) QVM *with* the VF approach. (b) SVM *with* the VF approach. (c) Log-based approach.

precision as NQ increases. The converged average PRs are $\mathrm{PR0} = 72.7\%$, $\mathrm{PR1} = 78.5\%$, and $\mathrm{PR2} = 79.5\%$.

We compare the proposed VF approach with an existing long-term learning method, that is, the log-based RF technique [25], which utilizes the user-specified relevant images as the seeds to search through the feedback logs and obtain more training samples for the SVM such that the average precision improves. Fig. 8c displays the average precisions as NQ increases. We observe that the average precisions also grow as the system accumulates more long-term logs. The converged average precisions are slightly lower than those obtained by using the SVM [15] with the VF method. It is because our VF approach can assist the system to significantly improve the average precisions at all displays; however, the log-based long-term learning is less helpful in improving the first display, since there is only one seed, namely, the user-submitted query image.

4.2.2 Retrieval Average Precisions at Different Feedback Iterations

Fig. 9a shows the in-breadth average PR performance with the NF using the QVM without the VF approach. The average PRs are plotted for various numbers of retrieved images presented at each iteration (in particular $r = 10$, 40, and 80). We observe that the average PRs converge to some upper bound as the NF increases. The average PRs obtained after five feedback iterations are 42.2 percent, 23.0 percent, and 17.5 percent for $r = 10$, 40, and 80, respectively. The in-breadth performance for the FRE without the VF approach is similar to that of QVM without the VF approach, and we omit it here for the lack of space. Fig. 9b shows the in-breadth analysis using the QVM with the VF approach. The VFs are obtained after experiencing 50,000 query sessions, as shown in Fig. 8a. The average PRs after five feedback iterations converge to 82.0 percent, 58.6 percent, and 57.7 percent for $r = 10$, 40, and 80, respectively, which are much greater as compared to those shown in Fig. 9a. The contribution in performance improvement takes place between PR0 and PR3. This is a desired property, since the users cannot stand too many feedback iterations, and they expect to retrieve greatly improved results after experiencing a very limited NF.

4.2.3 Virtual-Feature Values between Image Pairs

To realize the likelihood that two images with the same annotation receive a distance reduction through our VF mechanism, we compute the cosine values between the e-vector of each database image and that of all of its relevant images from the ground truth based on the learned VFs after experiencing 50,000 query sessions with 10 displayed images at every iteration. The percentages of cosine values, excluding cosine values of 0.0 and 1.0, are illustrated in Fig. 10. The percentages for cosine values of 0.0 and 1.0 are 97.3 percent and 1.4 percent, respectively. Note that the high percentage of 97.3 percent for cosine value 0.0 is due to the reason that our VF learned concepts are more specific than the true concepts (see Section 4.3). The sum of percentages with positive cosine values is 2.7 percent. This is the fraction of the data for which the reduction in distance takes place. If we take the average size of the image clusters, which is 180, for the UCR database, there are, on the average, $180 \times 2.7\% = 5$ relevant images that receive distance reduction and would very likely be retrieved in the first display. The number of relevant images that receive distance reduction in the second or later displays will

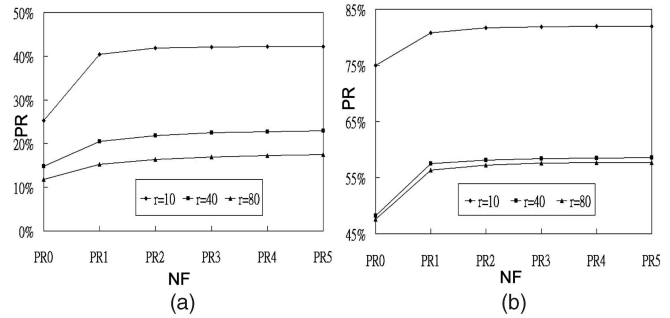


Fig. 9. Average PRs versus the NF. (a) Using the QVM without the VF approach. (b) Using the QVM with the VF approach.

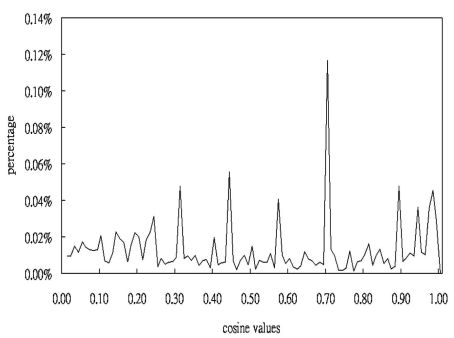


Fig. 10. Percentages of cosine values between relevant images.

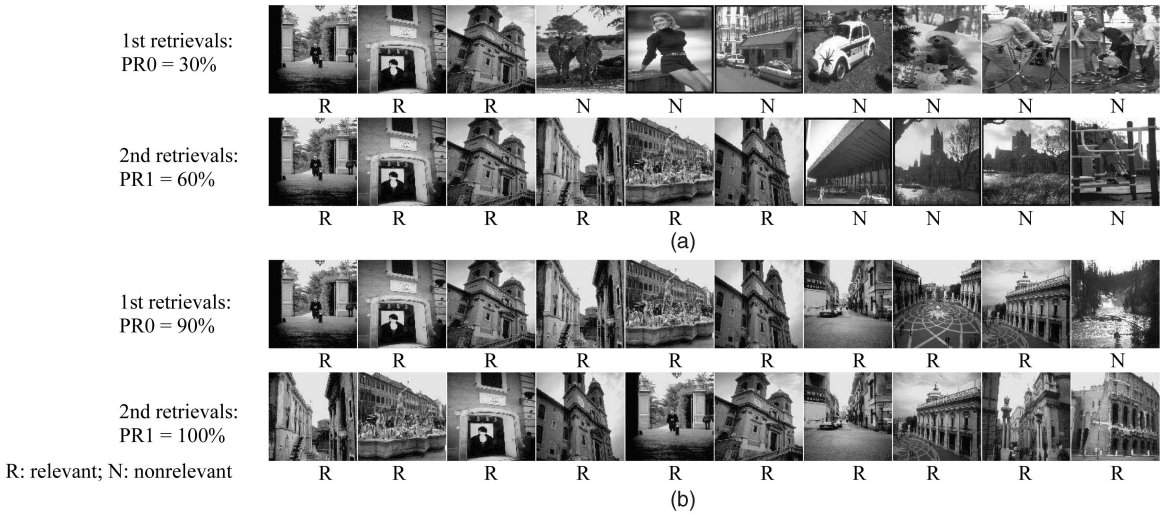


Fig. 11. Retrieval Example 1: Retrieval snapshots on the UCR database using the proposed QVM *with* the VF approach. (a) First instance of the query. (b) Second instance of the query.

significantly be higher than 5 due to the query formulation, which integrates the VFs of all previously retrieved relevant images as a new VF for the query and retrieve more relevant images.

Retrieval Example 1. Fig. 11 shows retrieval snapshots from our experiments by using the proposed QVM *with* the VF approach. Fig. 11a illustrates the query session when one building image is used for the first time as a query. Note that the query image is always the first image of each group (in Figs. 11, 12, and 19). In the first retrievals obtained before any feedback iteration, three images, including the query itself, are identified as relevant $(\mathrm{PR0} = 30\%)$. However, if we submit the same query image and use the euclidean matching for retrieval, the PR is only 10 percent. The reason that our VF mechanism is superior to the euclidean matching is due to the fact that relevant images have learned similar VFs when they were retrieved and identified as relevant in previous query sessions, thus resulting in the minimum value for our dissimilarity measure (5), even though they are not ranked in the top-10 closest images in the PF space. After the first feedback iteration, three additional relevant images are retrieved $(\mathrm{PR1} = 60\%)$ due to the query reformulation by computing the average PFs and the normalized VFs product from the relevant images. When the same building image is used for the second time as the query (Fig. 11b), the PR of the first retrievals is raised to 90 percent, since more database images have learned the VFs and propagated the VFs to other images. It is more likely to retrieve additional relevant images from the database that share some common VF concepts with the query. After the first feedback iteration, the query reformulation process contributes to search by one more relevant image, and the PR reaches 100 percent. It is noteworthy that the 10th relevant image in the last retrievals in Fig. 11b is ranked as the 414th closest image from the query in the PF space, and it is successfully retrieved in this query session by our VF approach using only two displays of retrievals. The next retrieval example shown in Fig. 12 is more significant in demonstrating this aspect, as will be noted. It is evident that our VF approach can effectively retrieve dissimilar images in the PF space that are semantically alike. This situation cannot be well handled by the QVM *without* the VF approach, since the QVM conducts short-term learning and forgets retrieval experiences in previous queries once a new query session is initiated.

Retrieval Example 2. The previous example demonstrates that the VF mechanism can be very helpful in improving retrieval performance when the user agrees with the query intention of mass users. On the other hand, when the user just partly agrees with the mass subject, the system will tune the weights between the long-term and short-term knowledge. Fig. 12 shows an example with a red-flower query image. The system retrieves 10 images that contain various kinds of flowers (see Fig. 12a) according to the VF statistics about the proportions of users who prefer to see flowers with different colors. However, if the current user is seeking for red flowers and marks only the first two images as relevant $(\mathrm{PR0} = 20\%)$, it will cause a precision degradation in the first retrieval. This feedback is very informative, and the embedded short-term RF technique (such as FRE) can increase the weight of color feature, since it is more important for discriminating relevant and nonrelevant objects under this scenario. Thus, the convergence of subsequent feedback process is expedited, although the precision obtained at the first retrieval could be worse than using the traditional technique such as the euclidean metric. For the retrieval system that uses the euclidean metric, the retrieved nonrelevant images are usually a mix of various kinds (for example, lions, skyscrapers, and sunsets), and they provide less information for the discrimination from the relevant images (with red flowers) and will prolong the feedback process. Our system realizes that the user just partly agrees with the mass subject, reduces the weight η for aged VF concepts, and assigns the relevant images with the new concept index. The VF of the query image formulated from the VFs of relevant images will also decrease the weight for aged VF concepts and increase the weight for the new concept. Thus, the system will become more like the embedded short-term RF technique (see (5)), since the e-vector of the query (contributed by the new concept) and the e-vector of the database images (focus on aged VF concepts) are almost orthogonal and cause near-zero cosine values. It is shown in Fig. 12a that after the first feedback iteration, the embedded QVM reformulates the query and retrieves five relevant images in the second retrievals $(\mathrm{PR1} = 50\%)$ and assigns

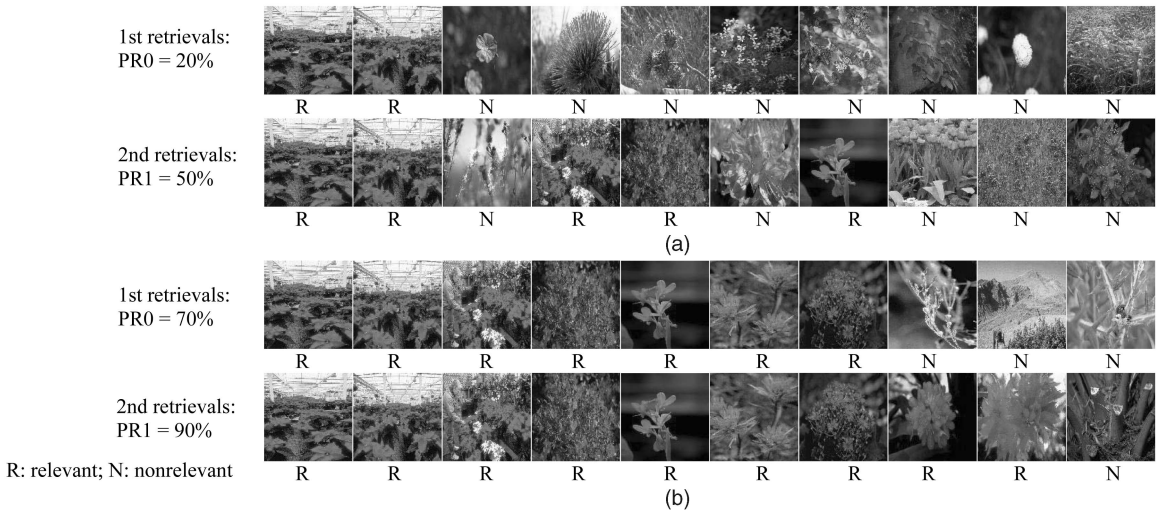


Fig. 12. Retrieval Example 2: Retrieval snapshots on the UCR database using the proposed QVM *with* the VF approach. (a) First instance of the query. (b) Second instance of the query.

them a new concept index. When the same red-flower image is used for the second time as the query (see Fig. 12b), the system retrieves seven relevant images in the first retrievals $(\mathrm{PR0} = 70\%)$. This is because more images have learned the "red-flower" VF concept between the two query sessions. The query is then reformulated by digesting the VFs of the seven relevant images and finds another two relevant images in the next retrievals $(\mathrm{PR1} = 90\%)$. The precision gain, as shown in Fig. 12b, is mainly due to the long-term learning conducted by the VF mechanism. The impact of this demonstrated example is more profound than that in Fig. 11. The ninth relevant image in the last retrievals in Fig. 12b is ranked as the 1,773rd closest image from the query in the PF space, and it is successfully retrieved in this query session by our VF approach. This is an extremely difficult problem for existing RF techniques to retrieve a semantically relevant image that is so far away from the query in the PF space using 10-image retrievals per display to conduct the RF.

4.3 Learning and Analysis

We assess the RA (6) between the human-labeled classes G and the partition matrix $M^{(v)}$ derived from MGCSs based on the VFs as the NQ v increases. Fig. 13 shows that the RA ratio rapidly grows as the NQ increases. This reveals that the retrieval performance obtained by the VF approach dramatically improves with the NQ. It is also observed that the larger the allowed number of the retrieved images r, the greater the RA ratio, and the ratio converges at a faster rate as well. This is because with a greater value of r, more relevant information can be learned at one feedback iteration. Considering the extreme case, if we allow the user to describe the relevant information for the entire database at every feedback iteration, the partition matrix $M^{(v)}$ will converge to G after only experiencing one query from each human-labeled class. Moreover, the case that the ground-truth G assigns two distinct images in different classes, whereas $M^{(v)}$ positions them in the same class never happened. This means that the learned concepts are actually a finer partition of the true concepts. In other words, the learned concepts are more specific than the true concepts, since the VF approach would relate two images with the same concept only if the two images are identified as relevant during the feedback.

4.3.1 Virtual-Feature Learning Rate

The VF learning conducted here is very different from the traditional offline learning in pattern recognition problems, where the training set is sequentially and disjointly retrieved, and the classification information for each training item is complete. Instead, our VF learning fits the real scenario of CBIR problems, and the training information is very limited due to the following reasons: 1) The learning task is online. The retrieval system conducts the user-feedback learning while many users are using it, so the training data are fed from the feedbacks in response to the retrievals of arbitrarily selected queries, which may not be the best candidates for the learner to improve its performance. Fig. 14 shows the number of distinct query images as the NQ increases. It is observed that only 6,324 distinct query images were selected as query images when the system has experienced 10,038 (the number of distinct images in the database) query sessions, and the system does not see all the distinct query images until it experiences more than 70,000 query sessions. 2) The training classification information is not complete for each query (only the relevant information in the display to this query is given), and the percentage of overlap between the training classification information attained at different query sessions is high. Let $Overlap^{(i)}$ be the percentage of relevant images in response to the ith query session that have

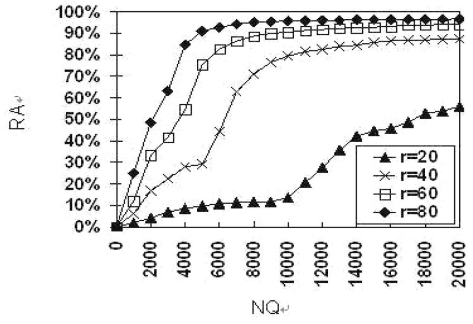


Fig. 13. RA ratio versus the NQ.

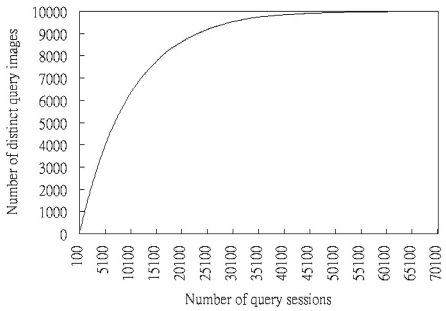


Fig. 14. Number of distinct query images versus NQ.

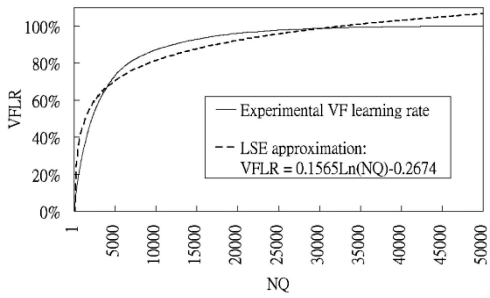


Fig. 16. VFLR curve and the LSE approximation.

previously been identified as relevant in another query session. Fig. 15 shows the curve for $Overlap^{(i)}$, where i denotes the NQ. We observe that the percentage of overlap between relevant images (the training information) increases with the NQ, which means that the amount of useful training information is limited. However, the *VFLR* is still efficient, as noted in the following.

The $VFLR$ can be defined as the fraction of the database images that have learned the VFs as the NQ increases. Let r be the number of displayed images at each iteration, $\delta^{(i)}$ be the percentage of displayed images that are identified as relevant for the ith query, and S be the number of images in the database. The $VFLR$ is computed as

$$VFLR = \sum_{i=1}^{NQ} r\delta^{(i)}\left(1 - Overlap^{(i)}\right)/S = \frac{r}{S}\sum_{i=1}^{NQ}\delta^{(i)}\left(1 - Overlap^{(i)}\right). \quad (7)$$

Since r/S is a constant, the growth rate of $VFLR$ is dependent on the variations of $\delta^{(i)}$ and $Overlap^{(i)}$ at each query session. For the best case, the retrieval system presents r truly relevant images at each query session ($\delta^{(i)} = 100\%$), and all those r relevant images have not previously been identified as relevant in another query session ($Overlap^{(i)} = 0\%$). Then, $VFLR = \frac{r}{S}\sum_{i=1}^{NQ} 1 = \frac{r}{S}NQ$, that is, $VFLR$ grows as a linear function of NQ. For the real case, we compute $VFLR$ that uses (7) by presenting query images chosen at random and calculating the values of $\delta^{(i)}$ and $Overlap^{(i)}$ from the retrievals. Fig. 16 shows the $VFLR$ curve with the NQ. The approximating logarithmic function of the $VFLR$ curve based on the least square error (LSE) criterion is derived as $VFLR = 0.1565Ln(NQ) - 0.2674$. The difference between the $VFLR$ analysis for the best and the real cases is due to the repeated selection of the same query images (see Fig. 14) and the overlap of relevant images in a query display that have been previously identified as relevant in another query display (see Fig. 15).

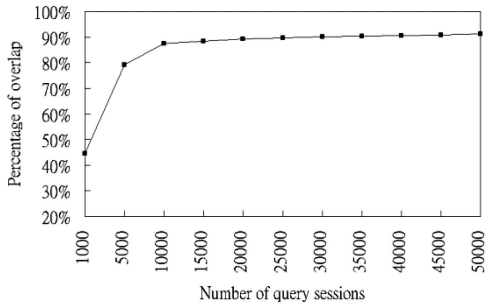


Fig. 15. Percentage of relevant images in the display in response to a query that have previously been identified as relevant in another query display.

4.3.2 Experiment on Visual Similarity

It should be noted that the VF approach not only partitions the images into classes based on human semantics but also presents the retrieved images according to visual similarity. We measure the normalized within-class visual variance between retrieved images at every feedback iteration j, which is defined as $S_W^j/n_+^j = \sum_{D_i \in R_j} \|D_i - \overline{D}\|^2 \big/ |R_j|$, where D_i is the feature vector of relevant images retrieved at the jth iteration, and $\overline{D}$ is the mean vector of these relevant images. We randomly select 100 query images and plot their average S_W^j/n_+^j in Fig. 17. We have the following observations. The retrievals at PR0 are based on the initial PFs and VFs of the original queries, so the images that are visually similar and are closely related in concepts to the original queries are first retrieved. These retrieved images would have a small within-class variance value. During the first two feedback iterations, the VFs of the already identified relevant images help in retrieving other relevant (but not so visually similar) images with common concepts. This is also revealed in Fig. 9b, where the major average precision improvement is achieved before PR2. The relevant images retrieved by our active nearest neighborhood are at a distance from each other, so the S_W^j/n_+^j curve initially rises from PR0 to PR2. After PR2, the query reformulation process

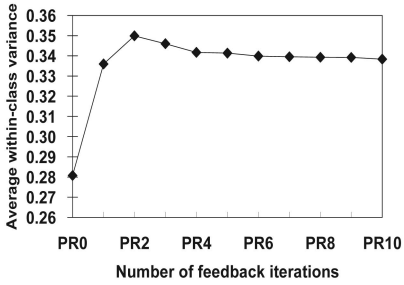


Fig. 17. Average within-class variance between retrieved relevant images versus the number of RF iterations.

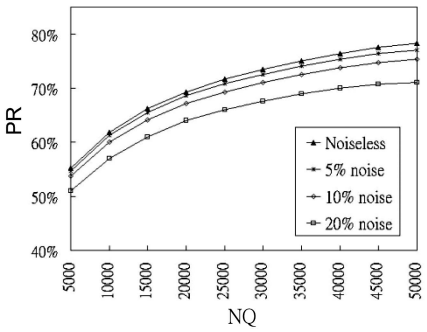


Fig. 18. Average PRs decrease with the amount of noise.

becomes relatively stable, since the changes in the identified relevant set are getting smaller, and very little new relevant information can be observed at this display. Thus, the contributions made by PFs are the main factor as reflected in S_W^t/n_+^t value that is decreased a little after PR2, and the newly retrieved relevant images (if any) at this display are more visually closer in the low-level feature space.

4.3.3 Noise Analysis

There are two types of noise when collecting relevant information. The first type of noise, referred as *false negative noise*, arises when a retrieved image is marked as nonrelevant when, in fact, it is truly relevant. On the other hand, the second type of noise referred as *false positive noise* is present when a nonrelevant image is retrieved, and it is mistakenly marked as relevant by the user. Considering the real-world CBIR scenarios, both types of noise are simultaneously added in equal proportions to the feedback. Various amounts of noise, in particular 5 percent, 10 percent, and 20 percent of the retrieved images, are randomly chosen and marked as noisy feedbacks, with equal amounts of false negatives and false positives. Fig. 18 shows that the average PRs obtained after two feedback iterations decrease with the amount of noise added to the feedback information. The decreased amount of PR, at the time of experiencing 50,000 query sessions, are 1.3 percent, 2.9 percent, and 7.2 percent for 5 percent, 10 percent, and 20 percent of noise, respectively.

4.4 Relevant Concept Transition

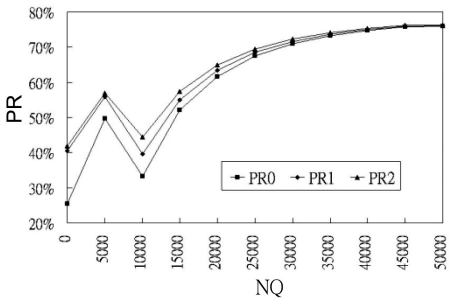
The proposed VF approach is self adapted in a relevance transitioning environment by tuning the relative importance between the aging concepts and the new concept (see (3) and (4)). All the results reported in Sections 4.2 and 4.3 are based on a static relevance ground truth, whereas the results presented in this section are based on two different ground truths. To simulate the relevance transitioning environment, one of the three persons involved in the original ground-truth generation (as described in Section 4.1) is asked to generate another ground truth for the database. This person relabels 600 images as follows: *First*, two original ground-truth classes (numbers 14 and 37) are removed, and the corresponding class members (a total of 452 images) are reassigned to the other 54 classes according to the intention of this person who is involved in the concept transition experiment. *Second*, from the original ground truth, 148 images with which this person had conflict in assigning class labels are chosen for relabeling by following this person's intent, and they are assigned to any of the 54 classes. Some sample images from the 600 relabeled images are shown in Fig. 19a.

4.4.1 Experiment on Concept Transition

The relevant concept transition experiment is conducted with 50,000 query sessions, and each query session goes through two RF iterations. From query sessions 1 to 5,000, the identification of relevant images is given according to the "original ground truth." From the 5,001st query session, the "new ground truth" takes over. Fig. 19b shows the variation of the average PRs with the NQ. It is observed that before the transition point, the VFs learn the relevant concepts of the original ground truth and help improve the retrieval performance. After the 5,000th query session, the performance is degraded with the new queries when the new ground truth is in conflict with the a priori relevant information learned earlier. The system automatically discounts the importance of the aging concepts and increases the weights for new concepts. From query sessions 5,001 to 50,000, the VFs learn the new ground truth, and the retrieval performance converges to a satisfactory level.

Retrieval examples. Fig. 20 illustrates some retrieval examples from the concept transition experiment, where the query image (image 144) is initially classified into class 15 (natural textured images) by the original ground truth and is then moved to class 2 (plant and flower images) after the 5001st query session by the new ground truth. When the 144th database image is used for the first time as a query at the 1,375th retrieval session (see Fig. 20a), a 70 percent PR has been reached due to the contribution of the learned VFs of relevant images in previous query sessions. The PR is then increased to 80 percent with the help of query reformulation. When this image is used for the second time

(a)



(b)

Fig. 19. The relevance concept transition experiment. (a) Sample images from the 600 relabeled images. (b) The average PRs versus the NQ.

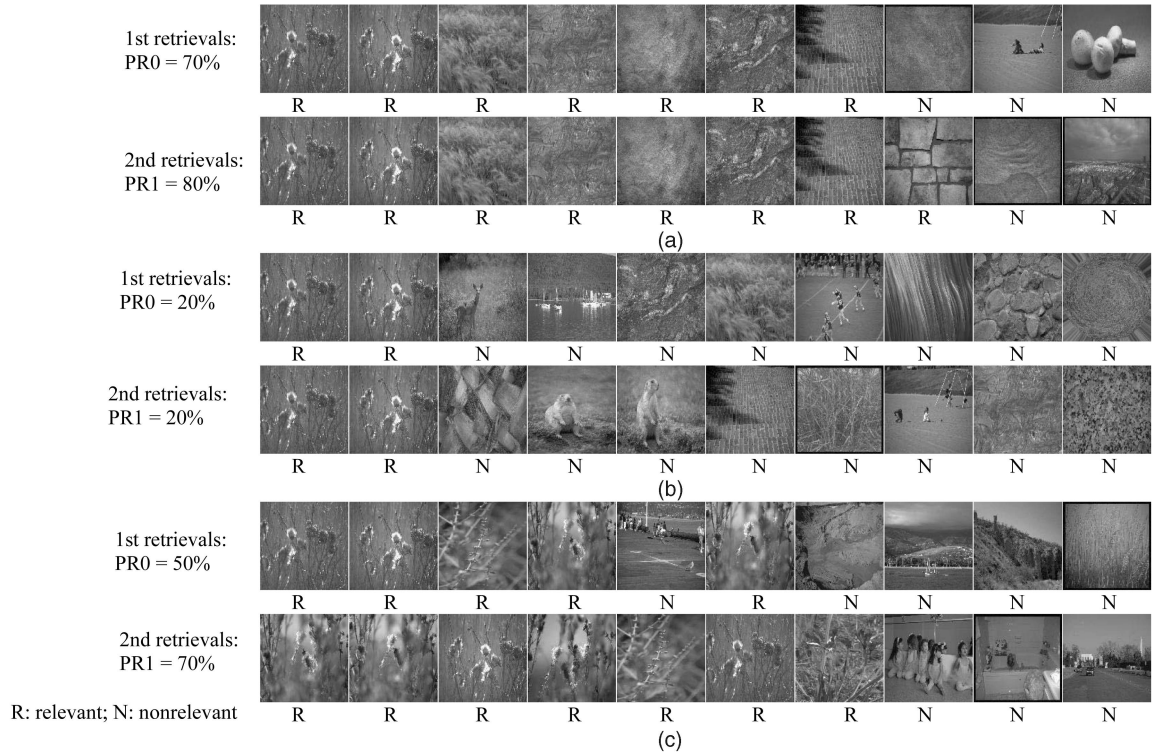


Fig. 20. Retrieval examples from the concept transitioning experiment using the proposed QVM with the VF approach. (a) First instance of the query: 1,375th query session. (b) Second instance of the query: 8519th query session. (c) Third instance of the query: 16,392nd query session.

as a query at the 8,519th query session (see Fig. 20b), the PR drops to only 20 percent, since the ground truth has been replaced by a new one, and the retrieved images that are relevant with the original ground truth are identified as nonrelevant with the new ground truth. Note that the second retrieved image is still identified as relevant, since it is also reclassified into class 2 by the new ground truth. With this significant drop (-60 percent) in the performance level, the system automatically discounts the aging VFs and increases the weight for new relevant concepts. We observe that the retrievals at the next iteration contain new images that do not share common VF concepts but are visually close to the query in the PF space, since the weight of the aging VFs has been discounted. When the 144th database image is, for the third time, used as a query at the 16,392nd retrieval session (see Fig. 20c), the PR is raised to 50 percent due to the increasing emphasis on the visual features. The PR is further improved to 70 percent by the query reformulation based on the learned VFs of relevant images. This example clearly demonstrates that the VFs can adaptively be learned in a relevant concept transitioning environment, and they can be managed to attain high PRs.

5 Conclusions

The traditional RF techniques used only within-session query experience to improve the retrieval precision. In this paper, we devised a new technique called the VF, which digests the cross-session query experiences to estimate the semantic relevance between images. Thus, the retrieval performance is improved by utilizing the a priori relevant information. Experimental results showed that the proposed retrieval system outperforms the one that applies the short-term RF alone. The concept learning conducted by the VF approach converges fast, so the retrieval performance of later query sessions can benefit from the VF-derived relevant information. To handle real-world dynamic database situation, the proposed method has a mechanism that changes the relevant concepts of a particular image by tuning the weights between the aging and the latest relevant information.

Compared with the existing RF and long-term learning techniques, the proposed method has the following salient features:

1. We assume neither the shape of the nearest neighborhood of the query nor the presence of one geometric cluster that contains all relevant images.
2. The relevant information of the original users' intention is directly stored in the VFs. This mechanism enables us to define an active (dynamically adapted) nearest neighborhood based on PFs and semantic features.
3. The proposed method combines the short-term and long-term relevance learning techniques to establish an effective retrieval system.
4. The proposed VF technique can be used for learning visual concepts involved in database images. It is also self adapted in a concept transitioning environment.
5. Our VF approach is a general framework that lets a short-term RF technique benefit from long-term learning.

References

[1] M. Flickner and H. Sawhney, "Query by Image and Video Content: The QBIC System," *Computer*, vol. 28, no. 9, pp. 23-32, 1995.

[2] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *Int'l J. Computer Vision,* vol. 18, no. 3, pp. 233-254, 1996.
[3] Y. Rui, T.S. Huang, S. Mehrotra, and M. Ortega, "Automatic Matching Tool Selection Using Relevance Feedback in MARS," *Proc. Second Int'l Conf. Visual Information Systems,* 1997.
[4] A. Yoshitaka and T. Ichikawa, "A Survey on Content-Based Retrieval for Multimedia Databases," *IEEE Trans. Knowledge and Data Eng.,* vol. 11, no. 1, pp. 81-93, Jan./Feb. 1999.
[5] A.W. Smeulders et al., "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.
[6] *Pattern Recognition,* special issue on image database, J.C.M. Lee and A.K. Jain, eds., vol. 30, no. 4, 1997.
[7] X.S. Zhou, Y. Rui, and T.S. Huang, *Exploration of Visual Data.* Kluwer Academic Publishers, 2003.
[8] J.J. Rocchio, Jr., "Relevance Feedback in Information Retrieval," *The SMART System,* G. Salton, ed., pp. 313-323, Prentice Hall, 1971.
[9] G. Ciocca and R. Schettini, "A Relevance Feedback Mechanism for Content-Based Image Retrieval," *Information Processing and Management,* vol. 35, no. 6, pp. 605-632, 1999.
[10] Y. Rui et al., "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 8, no. 5, pp. 644-655, 1998.
[11] J. Peng, B. Bhanu, and S. Qing, "Probabilistic Feature Relevance Learning for Content-Based Image Retrieval," *Computer Vision and Image Understanding,* vol. 75, nos. 1-2, pp. 150-164, 1999.
[12] B. Bhanu, J. Peng, and S. Qing, "Learning Feature Relevance and Similarity Metrics in Image Databases," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries (CBAIVL '98),* pp. 14-18, 1998.
[13] C. Meilhac and C. Nastar, "Relevance Feedback and Category Search in Image Database," *Proc. Int'l Conf. Multimedia Computing and Systems (ICMCS '99),* pp. 512-517, 1999.
[14] I. Cox et al., "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 20-37, 2000.
[15] S. Tong and E.Y. Chang, "Support Vector Machine Active Learning for Image Retrieval," *Proc. ACM Int'l Conf. Multimedia,* pp. 107-118, 2001.
[16] K. Tieu and P. Viola, "Boosting Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '00),* pp. 228-235, 2000.
[17] N. Vasconcelos and A. Lippman, "Learning from User Feedback in Image Retrieval Systems," *Proc. Neural Information Processing System,* 1999.
[18] A. Qamra, Y. Meng, and E.Y. Chang, "Enhanced Perceptual Distance Functions and Indexing for Image Replica Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 379-391, March 2005.
[19] T.P. Minka and R.W. Picard, "Interactive Learning with a 'Society of Models'," *Pattern Recognition,* vol. 30, no. 4, pp. 565-581, 1997.
[20] B. Bhanu and A. Dong, "Exploitation of Meta Knowledge for Learning Visual Concepts," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries (CBAIVL '01),* pp. 81-88, 2001.
[21] X. He, W.Y. Ma, O. King, M. Li, and H.J. Zhang, "Learning and Inferring a Semantic Space from User's Relevance Feedback for Image Retrieval," *Proc. ACM Int'l Conf. Multimedia,* pp. 343-346, 2002.
[22] F. Jing, M. Li, H.J. Zhang, and B. Zhang, "Relevance Feedback in Region-Based Image Retrieval," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 5, pp. 672-681, 2004.
[23] A. Dong and B. Bhanu, "Active Concept Learning in Image Databases," *IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics,* vol. 35, no. 3, pp. 450-466, 2005.
[24] P.Y. Yin, B. Bhanu, K.C. Chang, and A. Dong, "Integrating Relevance Feedback Techniques for Image Retrieval Using Reinforcement Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1536-1551, Oct. 2005.
[25] C.-H. Hoi and M.R. Lyu, "A Novel Log-Based Relevance Feedback Technique in Content-Based Image Retrieval," *Proc. ACM Int'l Conf. Multimedia,* pp. 24-31, 2004.
[26] Univ. of California, Riverside (UCR) Database, http://www.cris.ucr.edu/Database.html, Aug. 2007.
[27] K.R. Castleman, *Digital Image Processing.* Prentice Hall, 1996.

Peng-Yeng Yin received the BS, MS, and PhD degrees in computer science from the National Chiao Tung University, Hsinchu, Taiwan. From 1993 to 1994, he was a visiting scholar in the Department of Electrical Engineering, University of Maryland, College Park, and the Department of Radiology, Georgetown University, Washington D.C. In 2000, he was a visiting professor in the Visualization and Intelligent Systems Laboratory (VISLab), Department of Electrical Engineering, University of California, Riverside (UCR). From 2001 to 2003, he was a professor in the Department of Computer Science and Information Engineering, Ming Chuan University, Taoyuan, Taiwan. Since 2003, he has been a professor in the Department of Information Management, National Chi Nan University, Nantou, Taiwan, where he has been the chairman since 2004. He has been on the editorial board of the *International Journal of Advanced Robotic Systems*, the *Open Artificial Intelligence Journal*, *Open Artificial Intelligence Letters*, and *Open Artificial Intelligence Reviews* and has served as a program committee member and given invited talks at many international conferences. His current research interests include artificial intelligence, evolutionary computation, metaheuristics, pattern recognition, content-based image retrieval, relevance feedback, machine learning, computational intelligence, and computational biology. He has published more than 80 academic articles in reputable journals and conference proceedings. He is a member of the Phi Tau Phi Scholastic Honor Society. He received the Overseas Research Fellowship from the Ministry of Education in 1993, the Overseas Research Fellowship from the National Science Council in 2000, and the Best Paper Award from the Image Processing and Pattern Recognition Society of Taiwan. He is listed in *Who's Who in the World*, *Who"s Who in Science and Engineering*, and *Who's Who in Asia*.

Bir Bhanu received the SM and EE degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, the PhD degree in electrical engineering from the University of Southern California, Los Angeles, and the MBA degree from the University of California, Irvine. He is the founding professor of electrical engineering and served its first chair at the University of California, Riverside (UCR). Since 1991, he has been the cooperative professor of computer science and engineering and the director of the Visualization and Intelligent Systems Laboratory (VISLab). He is the founding director of an interdisciplinary Center for Research in Intelligent Systems (CRIS), UCR. He was a Senior Honeywell Fellow at Honeywell Inc., Minneapolis. He was with the faculty of the Department of Computer Science, University of Utah, Salt Lake City, Ford Aerospace and Communications Corp., California, INRIA, France, and IBM Research Laboratory, San Jose, California. He was the principal investigator of various programs for DARPA, NASA, NSF, AFOSR, ARO, ONR, and other agencies and industries in learning and vision, image understanding, pattern recognition, target recognition, biometrics, navigation, image databases, sensor networks, biologically inspired computation, and machine vision applications. He is the holder of 11 US and international patents. He was the general chair and program chair of several conferences and workshops. He has been on the editorial board of various journals and has edited special issues of several IEEE Transactions. He is a coeditor of *Computer Vision beyond the Visible Spectrum*, (Springer, 2004). He is a coauthor of *Computational Learning for Adaptive Computer Vision* (Springer, forthcoming), *Human Ear Recognition by Computer* (Springer, forthcoming), *Evolutionary Synthesis of Pattern Recognition Systems* (Springer, 2005), *Computational Algorithms for Fingerprint Recognition* (Kluwer, 2004), *Genetic Learning for Adaptive Image Segmentation* (Kluwer, 1994), and *Qualitative Motion Understanding* (Kluwer, 1992). He has published more than 250 reviewed technical publications in his areas of interest. He received two Outstanding Paper Awards from the Pattern Recognition Society and several industrial and university awards for research excellence, outstanding contributions, and team efforts. He is a fellow of the IEEE, the American Association for the Advancement of Science (AAAS), the International Association of Pattern Recognition (IAPR), and the International Society for Optical Engineering (SPIE).

Kuang-Cheng Chang received the BS degree in information management from Ming Chuan University, Taoyuan, Taiwan, in 2002 and the MBA degree in information management from the National Chi Nan University, Nantou, Taiwan, in 2007. His research interests include pattern recognition, machine learning, software engineering, and bioinformatics.

Anlei Dong received the BS and MS degrees in automation from the University of Science and Technology of China and the PhD degree in electrical engineering in 2004 from the University of California, Riverside (UCR). His research interests are machine learning and computer vision, including object detection, image retrieval, representation, understanding, indexing, and recognition.