JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 20, 1197-1212 (2004)

Short Paper

Learning Visual Concepts from Image Instances

JUN-WEI HSIEH, CHENG-CHIN CHIANG[1], YEA-SHUAN HUANG[2]
AND W. E. L. GRIMSON[3]
Department of Electrical Engineering
Yuan Ze University
Chungli, 320 Taiwan
E-mail: shieh@saturn.yzu.edu.tw
[1]*Department of Computer Science and Information Engineering*
National Dong Hwa University
Hualien, 974 Taiwan
[2]*Advanced Technology Center, Computer and Communication Laboratories*
Industrial Technology Research Institute
Hsinchu, 310 Taiwan
[3]*Artificial Intelligence Laboratory*
Massachusetts Institute of Technology
MA 02139-4307, U.S.A.

This paper presents a novel method of retrieving images by learning the commonality of instances from a set of training examples. The proposed scheme uses a coarse-to-fine algorithm to find the desired visual concepts from a set of instances for successful image retrieval. The learner at the coarse stage attempts to partition training data into two smaller compact sets (relevant and irrelevant) to reduce the size of the training examples, thus improving the efficiency of concept learning at the refined stage. At the refined stage, a proposed verification scheme is employed to verify each instance obtained at the coarse stage by examining its indexing and filtering capabilities based on a pool of images. Due to this extra examination step, the desired visual concepts can be learned more accurately, leading to significant improvement in image retrieval. Since no time-consuming optimization process is involved, all the desired visual concepts can be learned online. Experimental results are provided to verify the superiority of the proposed method.

Keywords: multiple instances, diverse density algorithm, relevance feedback, region instances, image retrieval

1. INTRODUCTION

With advances in content creation and World-Wide Web technologies, content-based multimedia information retrieval has received considerable attention in recent years

Received April 29, 2003; revised August 11, 2003 & September 19, 2003 & October 8, 2003; accepted November 21, 2003.
Communicated by Pau-Choo Chung.

[1-18]. This has led to the need for the development of content-based retrieval systems that will enable users to search for information directly via image contents. In the image-based retrieval systems developed so far, a commonly used approach is to exploit low-level features, such as colors [7-9], textures [10, 11], and shapes [12], as keys to retrieve images. Simplicity is the major benefit obtained from this approach. However, there still exists a great gap between low-level features and high-level concepts embedded in image contents. To address this challenge, many researchers [13-16] have investigated different learning methods for image retrieval. For example, Y. Rui *et al.* [13] proposed a re-weighting method to adjust the weight of a feature for the purpose of specifying its importance based on the user's relevance feedback. Picard and Minka [14] proposed a "FourEye" system to estimate the ideal query parameters using a "society of models." In addition, Maron and Ratan [16] described a diverse density algorithm to learn desired visual concepts from a set of positive and negative examples. Moreover, Tong and Chang [4] used techniques based on SVMs (support vector machines) and active learning to obtain useful relevance feedback for image retrieval. In their approach, the target concept is captured by a hyperplane in order to separate relevant images from irrelevant ones. This query refinement scheme can be regarded as a kind of pool-based active learning. At the beginning, the pool is the entire database of images, which is unlabeled. Then, according to the user's assignments, two sets of training samples, i.e., "relevant" or "irrelevant," are selected. Based on the training samples, the learner tries to find a desired concept by seeking the optimal hyperplane for effective separation of all the labeled images into different categories. However, this technique requires the use of a lot of training examples to train the desired visual concepts. In addition, the question of whether the learned visual concept has sufficient abilities to retrieve all the desired images from the database needs further study.

In this paper, a novel coarse-to-fine approach is employed to learn a set of important instances from a set of training samples for image retrieval. The coarse stage uses a kernel partition scheme to find a relevant class from a set of training examples. This stage can effectively reduce the amount of training data and, thus, speed up the efficiency of concept learning at the fine stage. This coarse stage achieves a similar goal of traditional clustering (or learning) schemes, like K means or SVMs, which try to find an optimal separation hyperplane for accurate data clustering. The fine stage tries to refine the obtained relevant class further by verifying its retrieval and filtering capabilities, which are important but are not considered in the above learning algorithms. In our proposed scheme, a pool of images is used to verify each instance of the found relevant class. The pool includes not only the user-provided training samples but also the results of the user's previous query. This extra consideration makes it possible to measure the retrieval and filtering capabilities of each instance. Since many redundant instances have already been filtered out through verification, the learned concept can be very accurately extracted for the purposed method of retrieving the desired images from the database. Furthermore, during the learning process, since no time-consuming steps are involved, all the visual concepts can be obtained online. Experimental results show that the proposed method indeed achieves great improvement in terms of retrieval accuracy, robustness, and stability.

The rest of the paper is organized as follows. In the next section, we first give an overview of relevance feedback learning. Then, the whole clustering procedure for learn-

ing the target concept is described in section 3. Section 4 provides details of the proposed verification algorithm. Section 5 describes the use of region instances in image retrieval. Section 6 reports experimental results, and conclusions are drawn in section 7.

2. LEARNING THE COMMONALITY OF INSTANCES BY MEANS OF RELEVANCE FEEDBACK

Relevance feedback (RF) learning is a widely accepted method for improving interactive retrieval effectiveness in image retrieval. The general scheme of RF learning is as follows: an initial search is made by the system with a user-provided query, and the system returns a small number of results to the user. In order to improve the retrieval accuracy, the user selects a set of positive and negative examples to retrain the desired visual concept by providing judgments of the results. Motivated by [16], we assume that a visual concept can be described by an instance or a set of instances, where an instance is a multidimensional feature vector used to describe different objects' properties like colors, shapes, textures, or geometries. Then, given several positive images $I_1^+, ..., I_{M_+}^+$ and negative images $I_1^-, ..., I_{M_-}^-$, we try to find an optimal concept t by maximizing the probability:

$$\Pr(t \mid I_1^+, ..., I_{M_+}^+, I_1^-, ..., I_{M_-}^-), \tag{1}$$

where M_+ and M_- are the number of positive and negative images, respectively, and t is a point in the feature space. Fig. 1 shows two types of instances used in this paper, i.e., the '+' and '□' types of instances shown in (a) and (b), respectively. Other types of instances can be found in [17]. Based on Eq. (1), various RF-related learning approaches have been proposed for finding the optimal solution t to maximize Eq. (1). For example, Maron *et al.* [16] introduced a diverse density algorithm to learn desired visual concepts from a set of positive and negative examples. In addition, Rui *et al.* [13] proposed a dynamical weight updating method to model high-level visual concepts with low-level features. If an instance is a region, the optimizing Eq. (1) can be interpreted as a way to find the best subimage t which is most consistent with all the regions of the training samples. However, a visual concept varies in appearance at different times and under different lighting conditions. For example, when a sky scene is described, it is clear that the color of the sky is close to "blue" or "white" on a sunny day; however, it will become "gray" or "dark" on a rainy day. Therefore, solution t used to optimize Eq. (1) cannot be only one single instance. Thus, instead of using one instance, we try to use a set of instances to represent a visual concept for better image retrieval. By rewriting Eq. (1), we try to find an optimal concept T by maximizing the probability:

$$\Pr(T \mid I_1^+, ..., I_{M_+}^+, I_1^-, ..., I_{M_-}^-), \tag{2}$$

where T is the conjunction of a set of instances and a set of weights.

Fig. 2 depicts our proposed learning algorithm. This is a coarse-to-fine approach to finding a set of dominant instances to bridge the gaps between low-level features and high-level knowledge. At the coarse stage, a proposed clustering technique is first

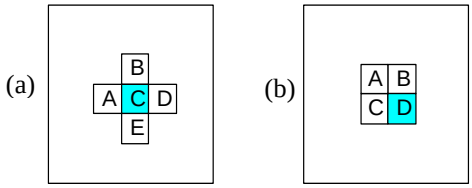


Fig. 1. Two types of instance features. (a) The '+' instance type, (b) The '□' instance type.

Proposed Algorithm

Input: user-given positive training samples $\{I_1^+, ..., I_{M_+}^+\}$ and negative ones $\{I_1^-, ..., I_{M_-}^-\}$.

Output: a set of dominant instances used to represent desired visual concepts.

Step 1: Feature extraction:

Extract two sets of instance features from $\{I_1^+, ..., I_{M_+}^+\}$ and $\{I_1^+, ..., I_{M_+}^+\}$, respectively.

Step 2: Hypothesis generation through instance partitioning:

A. Calculate the average distance between each instance and the set of positive images.

B. Calculate the average distance between each instance and the set of negative images.

C. Obtain the compact set C of instances to represent the desired visual concepts by means of kernel partitioning based on the distance calculations performed in Stages 2.A and 2.B.

Step 3: Concept refinement through instance verification:

A. Select the top-ranked and bottom-ranked images from the user's last query results as a verification pool.

B. Verify each instance in C according to the selected pool of images.

C. Obtain the refined visual concept $\overline{C}$ from C according to the verification process.

Fig. 2. Details of the proposed algorithm.

employed to partition the instance features into a smaller compact subset for speeding up the learning at the refined stage. At the refined stage, the obtained visual concept is further improved by verifying its indexing capability in retrieving all the desired images from the database. To do such verification, a pool of images is created to measure its varying retrieval and filtering capabilities. This paper applies a sampling method for selecting important images as the pool from the entire database to retrain the desired visual concepts. In what follows, details of the proposed clustering technique are described in section 3. Then, the verification algorithm is described in section 4.

3. LEARNING VISUAL CONCEPTS FROM INSTANCES BY CLUSTERING

As described above, in RF learning, a set of positive and negative training examples are selected in advance for retraining desired visual concepts. In this section, we propose a clustering technique for clustering the training samples into two different categories. The scheme tries to eliminate all possible irrelevant instances from the desired visual concept by calculating the distances between each instance and the positive and negative samples. Thus, a more compact subset of instances, which is closer to the positive samples but farther away from the negative ones, is generated. Details of the proposed clustering scheme are discussed in the following.

Let D_p and D_n be the sets of d-dimensional instances generated from the positive images $I_1^+, ..., I_{M_+}^+$ and from the negative images $I_1^-, ..., I_{M_-}^-$, respectively. In addition, let n_p and n_n be the numbers of elements in D_p and D_n, respectively. We can think of D_p and D_n as two hyper-balls enclosed within their centers by the radiuses $r_{p,\max}$ and $r_{n,\max}$, respectively. Here, $r_{p,\max}$ is the maximum relative distance between each instance in D_p and the other instances in the same ball. $r_{n,\max}$ is defined similarly for the set D_n. Let C_p and C_n be the kernels of D_p and D_n, respectively. If C is the concept we want to find, then the properties of C should be close to the kernel C_p but far from the C_n. Let the distance between two instances e_i and e_j be denoted as $\|e_i - e_j\|$ as follows:

$$\| e_i - e_j \| = \sum_k w_k (e_{ik} - e_{jk})^2, \tag{3}$$

where e_{ik} and e_{jk} are the k^{th} features of e_i and e_j, respectively, and w_k is a weighting factor for the k^{th} feature obtained from the inverse value of its corresponding variance. Let $dis(e_i, I_k)$ be denoted as the distance between e_i and an image I_k, i.e.,

$$dis(e_i, I_k) = \min_l \| e_i - e_l^{I_k} \|, \tag{4}$$

where $e_l^{I_k}$ is an instance in I_k. For an instance e_i in D_p, the distance between it and D_p is defined as

$$d(e_i, D_p) = \frac{1}{M_+} \sum_{k=1}^{M_+} dis(e_i, I_k^+). \tag{5}$$

Similarly, the distance between e_i and D_n is

$$d(e_i, D_n) = \frac{1}{M_-} \sum_{k=1}^{M_-} dis(e_i, I_k^-), \tag{6}$$

where I_k^- is one of the negative images. According to Eq. (5), the average value u_{d_p} of the distance $d(e_i, D_n)$ for all instances e_i in D_p can be defined by

$$u_{d_p} = \frac{1}{n_p} \sum_{e_i \in D_p} d(e_i, D_p). \tag{7}$$

Similarly, the average value u_{d_n} of $d(e_i, D_n)$ for all instances e_i in D_n can be calculated by

$$u_{d_n} = \frac{1}{n_n} \sum_{e_i \in D_n} d(e_i, D_n). \tag{8}$$

With u_{d_p} and u_{d_n}, we can define the kernels C_p and C_n for obtaining the desired visual concept C. In general, if an instance e_i belongs to kernel C_p or C_n, the distances between e_i and the sets D_p and D_n should be smaller than the average distances u_{d_p} and u_{d_n}, respectively. To preserve the representative capabilities of the set C, a larger radius $\hat{r}_p$ and a smaller $\hat{r}_n$ are used. Now, we define the radiuses $\hat{r}_p$ and $\hat{r}_n$ by setting $\hat{r}_p = 1.2u_{d_p}$ and $\hat{r}_n = 0.8u_{d_n}$, respectively. Then, for an instance e_i in D_p, we can determine whether e_i belongs to C_p according to the following rule:

$$e_i \in C_p \text{ if } d(e_i, D_p) < \hat{r}_p. \tag{9}$$

Similarly, the instance e_i belongs to C_n if $d(e_i, D_n) < \hat{r}_n$. Since C should be close to C_p but far from C_n, the desired concept C can be determined according to the following rule:

$e_i \in C$ if $e_i \in C_p$ and $e_i \notin C_n$ for each instance e_i in D_p.

When only one positive image is given, all the elements in C_p belong to C.

4. LEARNING VISUAL CONCEPT BY MEANS OF VERIFICATION

In the previous section, after clustering, we found a compact set C of instances to capture the user's query concepts. In what follows, a verification algorithm is proposed to refine C such that C has better abilities to retrieve all the desired images. As defined above, I^+ and I^- are the sets of user-provided positive and negative training images, respectively. In practice, if C is good enough, it will completely retrieve all the images in I^+ and filter out all images in I^- from the top list of retrieval results. Therefore, the basic idea behind the proposed verification algorithm is to verify whether each instance in C has the following properties or not:

(a) Let all the images in I^+ be ranked near the top of the retrieval list;
(b) Let all the images in I^- be ranked near the bottom of the retrieval list.

When the above idea is realized, a pool of images is needed to evaluate the retrieval and filtering capabilities of each instance in C.

Assume that the pool for instance verification includes K images. One straight-forward method is to randomly select K images from the entire database for such instance verification. However, the random selection scheme is ineffective. Here, we suggest that the training database can be selected from the user's previous query results. As illustrated in Fig. 3, the previous results should include many unknown training samples which are not returned to the user but can be selected in the pool for concept learning. The user applies a series of queries to retrieve the desired images from the pool of images.

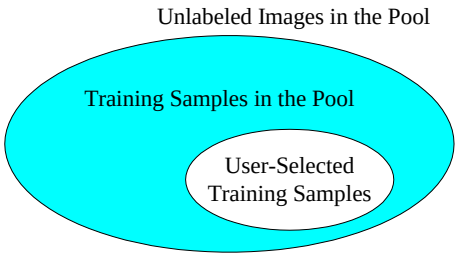
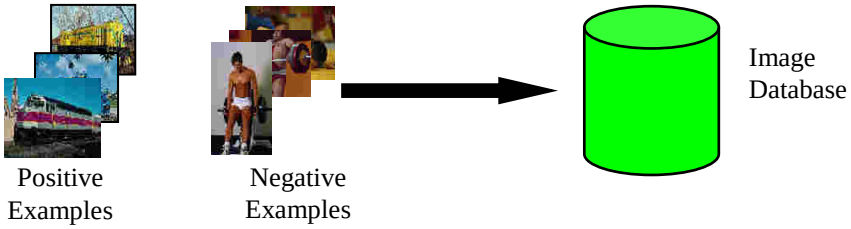
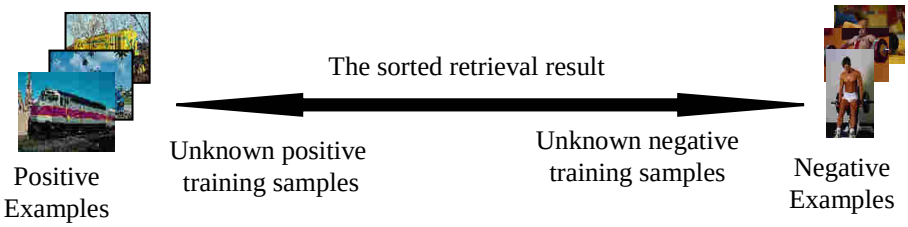


Fig. 3. Many training samples will not be return to and selected by the user.

As shown in Fig. 4, if one image in the entire database is close to the user's queries, it will be sorted close to the positive images but far from the negative images. Otherwise, it will be found at the bottom of the list. This situation indicates that most of the unknown positive images will appear at the top of the list of retrieval results, but that unknown negative images will be found at the bottom of this list. Based on this observation, we propose a new sampling strategy to create a new pool of training images: collect the top $0.5K$ retrieval results (denoted as $\bar{I}^{+}$) and the bottom $0.5K$ results (denoted as $\bar{I}^{-}$) as the set of unknown positive and negative samples, respectively. Then, the new pool of training samples is suggested as the union of $I^{+}, I^{-}, \bar{I}^{+}$, and $\bar{I}^{-}$, i.e., $Pool_{new} = \{I^{+}, I^{-}, \bar{I}^{+}, \bar{I}^{-}\}$. Therefore, the actual size of $Pool_{new}$ is $K + M^{+} + M^{-}$. In the following, a new method is proposed to verify each instance in C to ensure whether it is able to retrieve all the desired images from $Pool_{new}$.



(a) Querying by means of positive and negative images.



(b) Retrieval result.

Fig. 4. Possible locations of unknown training images in the retrieval result after querying. (a) Training samples before querying, (b) Retrieval result after querying.

Assume that e_i is an instance in C, and that S_{e_i} is the set of the top N query results obtained from $Pool_{new}$ when e_i is used as the query key. N is the number of retrieved images returned to the user. Given e_i, the distance between e_i and all the images in I^+ can be defined as

$$d(e_i, I^+) = \frac{1}{M^+} \sum_{I_k^+ \in I^+} dis(e_i, I_k^+) / MaxError, \tag{10}$$

where $dis(e_i, I_k^+)$ is defined in Eq. (4) and $MaxError$ is the maximum distance $dis(e_i, I_k^+)$ for normalization. In addition, according to $dis(e_i, I_k^+)$, each image I_k^+ in I^+ can be sorted and assigned using the sorted index $Index_{e_i}(I_k^+)$. If $index_{e_i}(I_k^+)$ is larger than N, image I_k^+ cannot be listed in S_{e_i}. Let $n_{e_i}^+(I^+)$ denote the number of images in I^+ which are listed in S_{e_i}. In addition, let $A_{e_i}^+(I^+)$ be the average of $Index_{e_i}(I_k^+)$ for all images in I^+, i.e.,

$$A_{e_i}^+(I^+) = \frac{1}{M^+} \sum_{I_k^+ \in I^+} Index_{e_i}(I_k^+). \tag{11}$$

Based on $d(e_i, I^+)$, $n_{e_i}^+$, and $A_{e_i}^+$, the retrieval ability of e_i is defined as

$$g_{e_i}(I^+) = n_{e_i}^+(I^+) + \exp\{-d(e_i, I^+) - A_{e_i}^+(I^+) / M_{Pool_{new}}\}, \tag{12}$$

where $M_{Pool_{new}}$ is the number of images in $Pool_{new}$. Similar to Eq. (12), we define $n_{e_i}^-(I^-)$ as the number of images in I^- which are not listed in S_{e_i}, and define $A_{e_i}^-(I^-)$ as the average of the inverse sorted indexes for all images in I^-, i.e.,

$$A_{e_i}^-(I^-) = \frac{1}{M^-} \sum_{I_k^- \in I^-} (M_{Pool_{new}} - Index_{e_i}(I_k^-)). \tag{13}$$

Next, we can measure the capability $f_{e_i}(I^-)$ of e_i to filter out the set I^- of negative images not listed in S_{e_i} as follows:

$$f_{e_i}(I^-) = n_{e_i}^-(I^-) + \exp\{d(e_i, I^-) - 1 - A_{e_i}^-(I^-) / M_{Pool_{new}}\}. \tag{14}$$

Let T_g be the average of g_{e_i} for all instances e_i if $n_{e_i}^+ > 1$, and let T_f be the average of f_{e_i} for all instances e_i if $n_{e_i}^- = M^-$. According to the retrieval and filtering capabilities g_{e_i} and f_{e_i} of e_i, we can determine whether e_i is a good candidate for representing the final desired visual concept. More precisely, if an instance e_i in C satisfies $g_{e_i} \geq T_g$ and $f_{e_i} \geq T_f$, then e_i is an element of $\bar{C}$, where $\bar{C}$ is the final set of instances used to represent the desired visual concepts. In practice, the value g_{e_i} is also a good confidence score for measuring the importance of e_i, i.e.,

$$w_{e_i} = g_{e_i} / \sum_{e_k \in \bar{C}} g_{e_k}. \tag{15}$$

Then, the distance between concept $\bar{C}$ and image I_k in the database can be calculated as follows:

$$d(\bar{C}, I_k) = \sum_{e_i \in \bar{C}} w_{e_i} dis(e_i, I_k). \quad (16)$$

With Eq. (16), each image can be sorted well and accurately retrieved from the database. In what follows, details of the proposed verification algorithm are presented.

Instance Verification Algorithm

Input: a visual concept C obtained from the kernel partition algorithm; N, the number of mages returned to the user; and Q, the user's last query result.

Output: the desired visual concept $\bar{C}$.

1. If Q is *NULL*, randomly select K images from the database, where $K = 8N$. Otherwise, select the top $0.5K$ images and the bottom $0.5K$ images from Q as training images. Let $Pool_{new}$ be the union of I^+, I^-, and the K selected images.
2. For each instance e_i in C,
   A. Calculate the $dis(e_i, I_k)$ according to Eq. (10), where I_k is an image in $Pool_{new}$.
   B. Sort all the images in $Pool_{new}$ in ascending order according to $dis(e_i, I_k)$. Accordingly, the sorted index, $Index_{e_i}(I_k)$ is obtained.
   C. Obtain the number $n_{e_i}^+(I^+)$ based on I^+ and $Index_{e_i}(I_k)$. Then, calculate the retrieval capability g_{e_i} of e_i according to Eq. (12).
   D. Obtain the number $n_{e_i}^-(I^-)$ based on I^- and $Index_{e_i}(I_k)$. Then, calculate the filtering capability f_{e_i} of e_i according to Eq. (14).
3. Obtain the average value T_g of g_{e_i} for all instances e_i if $n_{e_i}^+ > 1$. In addition to T_g, obtain the average T_f of f_{e_i} for all instances e_i if $n_{e_i}^- = M^-$.
4. For each e_i in C, if $g_{e_i} \geq T_g$ and $f_{e_i} \geq T_f$, then collect the e_i to the desired concept $\bar{C}$.

5. LEARNING VISUAL CONCEPT BY REGION INSTANCES

In Fig. 1, different instances are shown. These kinds of instances seem to be too low-level to represent image contents. Therefore, in this section, we will show that "regions" can also be useful instances for representing the desired image contents. Before using regions to represent image contents, each input image should first be segmented into pieces of regions. In this paper, the method proposed by Felzenszwalb and Huttenlocher [21] is adopted to roughly decompose an image into a set of regions.

In order to capture the characteristics of each segmented region, a set of features, including color, texture, and geometric features, are used as underlying primitives to represent the region. For the color feature in each region, four dominant colors are extracted using the K-means algorithm. Before clustering is performed, the initial values of the four dominant colors are obtained by dividing the region into four grids, and the average color of each grid is determined. After clustering is performed, each region can be represented by a set of color attribute pairs: $\{\{c_1, p_1\}, \{c_2, p_2\}, \{c_3, p_3\}, \{c_4, p_4\}\}$, where c_k is a 3-D color vector, p_k presents the weighting, and $\sum p_k = 1$. Given two regions e_i and

e_j whose color features are $e_i = \{\{c_{ik}, p_{ik}\}, k = 1, \ldots, 4\}$ and $e_j = \{\{c_{jk}, p_{jk}\}, k = 1, \ldots, 4\}$, respectively, the quadratic form is adopted to measure their distance, $d_C(e_i, e_j)$ [9].

In addition to the color feature, the texture features we use are the edge density and the histogram of edge orientations. Let $|e|$ be the area of a region e. Its edge density is defined by

$$E_e = \frac{\text{number of edge pixels in } e}{|e|}. \tag{17}$$

For the edge histogram, we divide edge orientations into 8 directions. Then, through counting the percentage of each direction for all the edge pixels in e, the so-called edge orientation histogram H_e can be obtained. Then, the texture distance between two regions e_i and e_j is

$$d_T(e_i, e_j) = 0.5 + \frac{|E_{e_i} - E_{e_j}|}{E_{e_i} + E_{e_j}} - \frac{1}{16}\sum_{k=0}^{7} \frac{\eta_k}{H_{e_i}(k) + H_{e_j}(k) - \eta_k}, \tag{18}$$

where $\eta_k = \min(H_{e_i}(k), H_{e_j}(k))$. In addition, the geometrical features we use are the area, the eccentricity, and the orientation of a region e. Before all the geometrical features are extracted, each image is first normalized to a standard size of 256×256. The eccentricity R_e is the ratio of the major axis to the minor axis of region e. The orientation of the region can be decided based on the central moments of e as follows:

$$\alpha_e = \frac{1}{2}\tan^{-1}\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}, \tag{19}$$

where $\mu_{p,q}$ is the $(p + q)$th central moment of e. Then, the distance between geometric features of e_i and e_j is defined by

$$d_g(e_i, e_j) = \frac{2}{3} \times \left(\frac{||e_i| - |e_j||}{|e_i| + |e_j|} + \frac{|R_{e_i} - R_{e_j}|}{R_{e_i} + R_{e_j}} + \frac{|\alpha_{e_i} - \alpha_{e_j}|}{2\pi}\right). \tag{20}$$

In this paper, the image database prepared for processing focuses on the domain of natural images. The color feature is considered to be more important than the other features for retrieving images. Then, the integrated distance between e_i and e_j is defined as follows:

$$||e_i - e_j|| = w_c d_c(e_i, e_j) + w_t d_t(e_i, e_j) + w_g d_g(e_i, e_j), \tag{21}$$

where $w_c = 0.8$, $w_t = 0.1$, and $w_g = 0.1$. In Eq. (3), we have defined the distance between any two instances. If the instance type is "regions," Eq. (21) is adopted to replace Eq. (3) for distance calculation. Other formulations in our proposed algorithm remain the same.

6. EXPERIMENTAL RESULTS

In order to analyze the performance of our algorithm, an interactive image retrieval system which allows users to search for their desired images from a database was implemented. The database contained 3000 images which came from 30 categories of COREL photo galleries. The categories included various visual classes like cars, human, waterfalls, mountains, sunsets, and so on. During feature extraction, for the instance feature, each image was first smoothed using a Gaussian filter and then down-sampled to obtain an 8×8 image, from which 36 instances were generated to represent the contents of this image. Each instance feature was defined by the mean RGB values of the central pixel and the relative color differences with its four neighbors. The RGB values were normalized to be in the $[0, 1]^3$ cube. Therefore, this instance type was represented as a vector with 15 elements. For the region instance, each image was smoothed first, using a Gaussian filter, and then normalized to obtain a 256×256 image. Then, each image was segmented into several regions, from which the 10 largest regions were selected to represent the desired image contents.

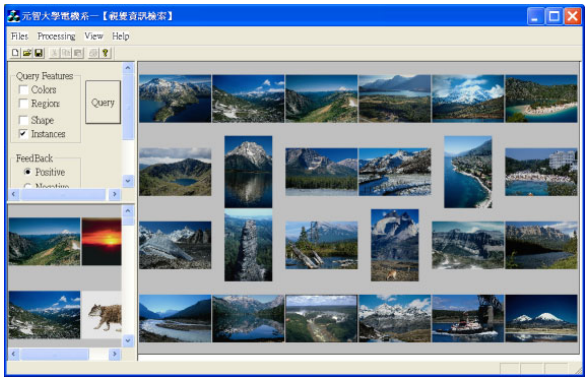
In our experiments, the color histogram approach [7], the diverse density (DD) algorithm [17] with the noisy-or model, the composite region templates (CRT) approach [20], and the integrated region matching (IRM) approach [19] were implemented for the purpose of comparison. For the diverse density algorithm, according to the suggestion made in [17], only the best concept point was used to represent the visual concept. With the histogram approach, each color channel (R, G, and B) was quantized into 8 levels such that a 512-dimentional color histogram feature vector was extracted from each image. As for other existing region-based retrieval algorithms, such as the NeTra system [22] and the Blobworld system [18], these methods need additional information to make additional judgments in the selection of important query regions. Therefore, we compared our method with these two methods by assuming that the largest region in the query image was the most important query region for retrieving images.

Initially, this system allows users to query images using features such as colors, shapes, textures, or instances. Fig. 5 (a) shows a snapshot of the system in action when used to retrieve the sunset concept with '+' instances using three positive training samples and two negative ones. Fig. 5 (b) shows another snapshot of the system when used to retrieve the mountain concept with the same type of instance features but based on four positive instances and two negative ones. To evaluate the performance of the proposed method, the precision and recall graphs were used to measure the retrieval performance. Precision is the ratio of the number of correct images to the number of retrieved images. Recall is the ratio of the number of correct images to the total number of correct images in the database. The two measures are defined as follows

$$Precision(N) = C_N / N \text{ and } Recall(N) = C_N / M,$$

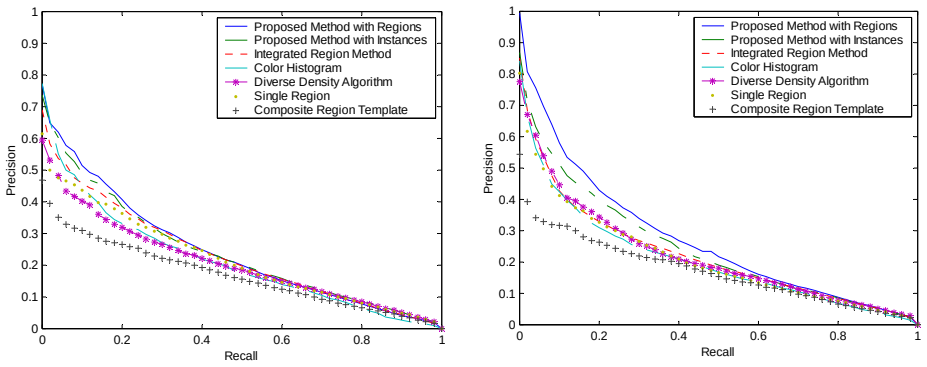
where N is the number of retrievals, C_N the number of relevant matches among all the N retrievals, and M the total number of relevant matches in the database obtained through a subjective testing.

(a)



(b)

Fig. 5. Retrieval results obtained using the proposed method. (a) Results of the sunset concept when three positive training images and two negative ones were given, (b) Results of the mountain concept when four positive training images and two negative ones were given.



(a) (b)

Fig. 6. Performance comparison for different approaches. (a) Only one query image was used. (b) Two or three query images were used. Clearly, the proposed method performed the best.

Here, we use two kinds of instances to demonstrate the learning capabilities of our proposed method. The first one is the '+' type of instance shown in Fig. 1 (a), and the second one is the 'region' instance, which has been discussed in section 5. Fig. 6 (a) shows a comparison of the performance achieved using the proposed method with the '+' instance type, the proposed method with region instances, the DD algorithm [17], and the color histogram approach [7], the IRM approach [19], the CRT method [20], and the NeTra-like system [22] when only one positive image was given. During testing, we created a potential training set that consisted of more than 100 images randomly chosen from each category. Each image in this set was chosen as a seed to query different image contents. From this figure, it can be clearly seen that the proposed method with regions performed better than the other methods. Since the '+' instance feature is lower-level than the region feature, our method with the '+' instance performed worse than the 'region instances' method but still outperformed the other methods. This figure also proves that the region-based methods (like the IRM approach and the NeTra-like system) performed better than the other low-level features (like the color histogram). Among these region-based methods, our proposed method can automatically analyze important regions and their associated weights for the purpose of accurate content representation. Thus, better retrieval results can be obtained using our proposed method. As for the DD algorithm, since only one query image is given, it cannot have enough samples to accurately learn the desired visual concepts. Therefore, other methods like the IRM approach, color histogram, and the NeTra-like method work better than the DD algorithm. As for the CRT approach, since it concerns only spatial structures, it works well only when spatial relations are important in content descriptions.

In the previous experiment, only one image was used to train the desired visual concept. Due to great variety of images, certainly, a single query image cannot provide enough information for concept learning and content description. Therefore, another experiment was performed to demonstrate the learning effects of our method when more query images are provided. In this experiment, a potential training set was created by picking two or three positive and negative examples for learning a given concept. With the histogram-based approach, the CRT approach, and the IRM approach, their performances was measured by averaging the total retrieval performance of each image in one query set. Fig. 6 (b) shows a comparison of the performance of the different retrieval schemes when two or three examples were given. Due to the learning capabilities, the DD algorithm performed better than the color histogram, the CRT method, and the NeTra-like method. Our proposed method still performed the best since its learning capabilities can filter out many redundant instances in advance. In the experiment, our proposed method used only 0.1 seconds on average to learn the desired visual concepts. Based on the above experimental results, the superiority of the proposed method has been verified.

7. CONCLUSIONS

In this paper, we have presented a coarse-to-fine method to efficiently extract the commonality of instances from a set of training samples for image indexing. This method first partitions training samples into a relevant class and another irrelevant class, and then

refines the extracted relevant class more accurately using a verification process. The clustering scheme is applied to speed up concept learning at the fine stage. The verification process verifies each remaining instance (after the above clustering process is performed) by calculating the retrieval and filtering capabilities of the instance. This verification process is very different from other learning methods which consider "learning" based only on user-selected training samples and do not check the above retrieval and filtering capabilities of the selected concept. Experimental results showed that our method performed extremely well in terms of retrieval accuracy, efficiency, and stability.

ACKNOWLEDGMENT

This paper is a partial result of the project No. A311XS1213 conducted by the ITRI under sponsorship of the Minister of Economic Affairs, Taiwan, R.O.C.

REFERENCES

1. J. Smith and S. Chang, "VisualSEEK: a fully automated content-based image query system," in *Proceedings of ACM International Conference on Multimedia*, 1996, pp. 87-98.
2. Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: current technique, promising direction, and open issues," *Journal of Visual Communication and Image Representation*, Vol. 10, 1999, pp. 39-62.
3. W. Niblack *et al.*, "The QBIC project: querying images by content using color, texture, and shape," in *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, Vol. 1908, 1993, pp. 173-187.
4. S. Tong and E. Chang, "Support vector machine active learning for image retrieval," *ACM International Conference on Multimedia*, 2001, pp. 107-118.
5. J. R. Bach *et al.*, "The virage image search engine: an open framework for image management," in *Proceedings of SPIE and Retrieval for Image and Video Databases*, Vol. 2670, 1996, pp. 76-87.
6. A. Pentland, R. W. Oicard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," *International Journal of Computer Vision*, Vol. 18, 1996, pp. 233-254.
7. M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, Vol. 7, 1991, pp. 11-32.
8. J. Huang, S. Mehrotra, M. Mitra, W. J. Zhu, and R. Zabih, "Image indexing using color correlogram," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 762-768.
9. J. Hafner *et al.*, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, 1995, pp. 729-736.
10. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Transactions on Systems Man and Cybernetics*, Vol. 3, 1973, pp. 610-621.

11. B. S. Manjunath and M. Y. Ma, "Texture feature for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, 1996, pp. 837-842.
12. A. K. Jain and A. Vailaya, "Shape-based retrieval: a case study with trademark databases," *Pattern Recognition*, Vol. 31, 1998, pp. 1369-1390.
13. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8, 1998, pp. 644-655.
14. T. P. Minka and R. W. Picard, "Interactive learning using a society of models," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 1996, pp. 447-452.
15. P. Lipson, E. Grimson, and P. Sinha, "Context and configuration based scene classification," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 1997, pp. 1007-1013.
16. O. Maron and A. L. Ratan, "Multiple instance learning from natural scene classification," in *Proceedings of 14th International Conference on Machine Learning*, 1997, pp. 341-349.
17. A. L. Ratan, "Learning visual representation for classification," Doctoral Thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., 1999.
18. S. Belongie, C. Carson, H. Greenspan, and J. Malick, "Color and texture based image segmentation using EM and its application to content-based image retrieval," *IEEE International Conference on Computer Vision*, 1998, pp. 675-682.
19. J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantic-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 947-964.
20. J. R. Smith and C. S. Li, "Image classification and querying using composite region templates," *Computer Vision and Image Understanding*, Vol. 75, 1999, pp. 165-174.
21. P. F. Felzenszwalb and D. P. Huttenlocher, "Image segmentation using local variation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 1998, pp. 98-104.
22. Y. Deng and B. S. Manjunath, "An efficient low-dimensional color indexing scheme for region-based image retrieval," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Vol. 6, 1999, pp. 3017-3020.
23. R. G. Gonazlez and R. E. Woods, *Digital Image Processing*, Addison Wesley, 1992.

Jun Wei Hsieh (謝君偉) received his Ph.D. degree in Computer Engineering from the National Central University, Taiwan, in 1995. He got the Phai-Tao-Phai award when he graduated. From 1996 to 2000, he was a Researcher Fellow at the Industrial Technology Researcher Institute, Hsinchu, Taiwan, and managed a team to develop video-related technologies. He is presently an Associated Professor at the Department of Electrical Engineering, Yuan Ze University of Taiwan. His research interests include content-based multimedia databases, video indexing and retrieval, computer vision, and pattern recognition.

Cheng-Chin Chiang (江政欽) received his Ph.D. degree in Computer Science and Information Engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C. in 1993. From 1993 to 2000, he was a researcher of Advanced Technology Center of Computer and Communication Research Laboratories in Industrial Technology Research Institute (ITRI), Hsinchu, Taiwan. In August 2000, he joined the faculty of the Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan. His research interests include neural networks, pattern recognition, multimedia systems, virtual reality, and content-based multimedia retrieval. Dr. Chiang received the honor of Long-term Doctorial Thesis Award from Acer Corporation in 1993, the honor of Outstanding Research Achievements Award form ITRI in 1996, the honor of Outstanding Young Engineer Award from Chinese Institute of Engineers in 2000.

Yea-Shuan Huang (黃雅軒) graduated from the Computer Science Department of Concordia University, Canada in 1994. Currently, he is a senior researcher of Advanced Technology Center of Computer and Communication Research Laboratories in Industrial Technology Research Institute (ITRI), Hsinchu, Taiwan, and he is the leader of the Interactive Visual Information Processing project. His research interests include image processing, pattern recognition, and neural networks.

Eric Grimson is a Professor of Computer Science and Engineering at the Massachusetts Institute of Technology, and holds the Bernard Gordon Chair of Medical Engineering. He also holds a joint appointment as a Lecturer on Radiology at Harvard Medical School and at Brigham and Women's Hospital. He received a B.S. (Hons) in Mathematics and Physics from the University of Regina in 1975 and a Ph.D. in Mathematics from MIT in 1980. Prof. Grimson currently heads the Computer Vision Group of MIT's Artificial Intelligence Laboratory, which has pioneered state of the art systems for object recognition, image database indexing, image guided surgery, target recognition, site modeling and many other areas of computer vision. Recently, his group has been active in applying vision techniques in medicine: for image guided surgery, minimally invasive surgery and telemedicine.