Mapping Low-Level Features to High-Level Semantic Concepts in Region-Based Image Retrieval*

Wei Jiang
Department of Automation
Tsinghua University
Beijing 100084, China

Kap Luk Chan
School of E.E.E.
Nanyang Technology University
Singapore, 639798

Mingjing Li Hongjiang Zhang
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China

Abstract

In this paper, a novel offline supervised learning method is proposed to map low-level visual features to high-level semantic concepts for region-based image retrieval. The contributions of this paper lie in threefolds. (1) For each semantic concept, a set of low-level tokens are extracted from the segmented regions of training images. Those tokens capture the representative information for describing the semantic meaning of that concept; (2) A set of posteriors are generated based on the low-level tokens through pairwise classification, which denote the probabilities of images belonging to the semantic concepts. The posteriors are treated as high-level features that connect images with high-level semantic concepts. Long-term relevance feedback learning is incorporated to provide the supervisory information needed in the above offline learning process, including the concept information and the relevant training set for each concept; (3) An integrated algorithm is implemented to combine two kinds of information for retrieval: the information from the offline feature-to-concept mapping process and the high-level semantic information from the long-tern learned memory. Experimental evaluation on 10,000 images proves the effectiveness of our method.

1. Introduction

Content-based image retrieval (CBIR) has been extensively explored since last decade. It is well known that the retrieval performance of CBIR systems is hindered by the gap between high-level semantic concepts and low-level visual features [11]. *Relevance Feedback* and *Region-Based Image Retrieval (RBIR)* have been proposed as two promising solutions to bridge this gap. Relevance feedback approaches can be classified into two categories: the *short-term relevance feedback learning (SRF)* is generally treated as the online supervised learning process [10, 12, 14], where the user labels some images to be "relevant" or "irrelevant" to his query concept during each feedback round and helps the system to successively refine query and give better retrieval results in the next feedback round; the *long-term relevance feedback learning* (*LRF*) [3, 4, 5] memorizes the information labeled by the user during the feedback process, and accumulates the information as semantic experience to help retrieval in subsequent query sessions. RBIR approaches segment images into several regions and use region-based low-level features, which represent images at the object level, to retrieve images instead of global features of the entire image. Most previous RBIR methods directly calculate the region-to-region or the image-to-image similarity for retrieval [1, 2, 13]. Recently, relevance feedback learning is introduced into RBIR by [6, 15].

To bridge the feature-to-concept gap, CBIR systems should solve two problems: *feature extraction* and *feature selection*. Other than conventional low-level visual features (e.g. color histogram or wavelet texture), new features should be learned, which are more representative to describe the semantic meaning of concepts. Also, for a specific concept, we should tell that which features are more representative than the others. In the CBIR context, feature extraction and feature selection can be carried out in two ways. (1) During the online SRF process: Most previous systems carry out online feature extraction or selection using the labeled images from the user as training samples. The Boosting feature selection method [12] is a typical approach. Since the labeled training samples are usually very few compared with the feature dimensionality and the size of the database, the performance is often unsatisfactory. (2) During an offline learning process: More training samples may be obtained, and better performance can be expected for the offline feature extraction and feature selection. As far as the authors know, previously reported works seldom address on this issue. The latest RBIR approaches in [6, 15] treat offline feature extraction in the form of unsupervised learning. They learn new features out of the region pool

*The work was performed at Microsoft Research Asia.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00© 2005 IEEE

from all images in the database, and represent images by the extracted features instead of by conventional region-based low-level features for retrieval. However, feature selection is not carried out in the unsupervised approaches, whose effectiveness is limited.

To conduct offline feature selection, the supervisory information about high-level semantic concepts is needed. Intuitively, this kind of information can be obtained from LRF learning, because the long-term cumulate memory provides high-level semantic information, e.g., some images come from the same semantic concept and some images should be classified into different concepts. Our recent work in [5] proposes an LRF learning mechanism, which learns real-world semantic concepts underlying images and records positive training samples for each learned concept. The supervisory information in the cumulate experience can be exploited in the learning process of the offline feature extraction and feature selection (see Sec.3 for details).

The contributions of this paper are mainly in three folds. (1) Offline feature extraction and feature selection: For each semantic concept, a set of new features – ***low-level tokens*** – are extracted based on images in this concept. For every individual concept, the extracted low-level tokens are specifically learned according to this concept, and they contain the most representative information in describing the semantic meaning of this concept. Thus through extracting low-level tokens, feature extraction and feature selection are simultaneously accomplished. (2) A set of posteriors are generated through pairwise classification based on low-level tokens, which denote the probabilities of images belonging to the semantic concepts. The posteriors are treated as high-level features which connect images to high-level semantic concepts. The above two steps together are called the process of ***mapping low-level features to high-level semantic concepts (feature-to-conceptmapping),*** because the posterior-based features are generated from region-based low-level features through extracting low-level tokens. This mapping process is carried out offline in the form of supervised learning. LRF learning is adopted to provide the supervisory information – information about semantic concepts and the "relevant" training samples for each concept. (3) The offline learned posterior-based features and the high-level semantic information from LRF records are incorporated to retrieve images. SRF is carried out based on the posterior-based features. LRF is carried out after every query session when users use our system to retrieve images. The system evolves through time. When more and more retrievals are conducted, the semantic knowledge keeps growing, and the posterior-based features map low-level features to high-level semantic concepts better and better. We evaluate our algorithm and compare it with the state-of-the-arts based on 10,000 images. Experimental results show that out method can effectively improve the retrieval performance of the RBIR system.

2. Feature-to-Concept Mapping

Fig. 1 illustrates the entire process of mapping low-level features to high-level concepts through extracting low-level tokens. We will describe the detailed steps in this part.

2.1. Low-level features & image segmentation

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be N images in a database. An image $\mathbf{x}_i$ is firstly partitioned into small tiles without overlapping, and feature vectors (color features and texture features) are extracted based on each tile. Then $\mathbf{x}_i$ is segmented into a set of regions $\Phi(\mathbf{x}_i) = \{r_1(\mathbf{x}_i), \ldots, r_{m_i}(\mathbf{x}_i)\}$ through the spectral clustering method proposed in [16]. Each region is represented by the mean feature vectors of the member tiles in it, and the regions correspond to a set of saliency memberships $\mathcal{V}(\mathbf{x}_i) = \{v_1(\mathbf{x}_i), \ldots, v_{m_i}(\mathbf{x}_i)\}$ which describe how well the regions represent the image. $\mathcal{V}(\mathbf{x}_i)$ can be determined by many mechanisms, such as by the region size or position in the image. In our system $\mathcal{V}(\mathbf{x}_i)$ is measured by the ***computational visual attention.*** The static attention model described in [8] is adopted to calculate the saliency map of the entire image, and $v_j(\mathbf{x}_i)$ is given by the average saliency value of the member pixels in $r_j(\mathbf{x}_i)$.

2.2. Extracting low-level tokens

Let $S_1, \ldots, S_M$ be M semantic concepts, the system extracts different low-level tokens, respectively, for different concepts. Let $\mathcal{X}^i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{N_i}^i\}$ be the set of images belonging to the concept S_i. $\Re^i = \bigcup_{j=1}^{N_i} \Phi(\mathbf{x}_j^i)$ is all the segmented regions from all the images in S_i. The ***Principle Component Analysis (PCA)*** is exploited based on $\Re^i$ to reduce the dimensionality of the original feature space, which reduces the redundancy of the image representation by low-level region-based features, and alleviates the singularity problem in matrix computation during the token extraction process. After PCA, the original $\Re^i$ is projected to be $\tilde{\Re}^i$. Rename elements in $\tilde{\Re}^i$ as $\tilde{\Re}^i = \{r_1^i, \ldots, r_m^i\}$. We assume that region r_j^i is generatively formed by n_i low-level tokens $C_1^i, \ldots, C_{n_i}^i$ through a ***Gaussian Mixture Model*** (***GMM***) as:

$$p(r_j^i) = \sum_{k=1}^{n_i} p(C_k^i) p(r_j^i | C_k^i) \tag{1}$$

where $p(r_j^i | C_k^i)$ follows the Gaussian distribution:

$$p(r_j^i | C_k^i) = (2\pi |\Sigma_k^i|)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2}(r_j^i - \overline{\mu}_k^i)^T (\Sigma_k^i)^{-1} (r_j^i - \overline{\mu}_k^i)\right\}$$

$\overline{\mu}_k^i$ and Σ_k^i are the mean vector and covariance matrix of the token C_k^i respectively. The mixture-based grouping method with model selection proposed in [7] is adopted to extract $C_1^i, \ldots, C_{n_i}^i$ by clustering the regions in $\tilde{\Re}^i$. This method

Proceedings of the 2005 **IEEE Computer Society Conference on Computer Vision and Pattern Recognition** (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

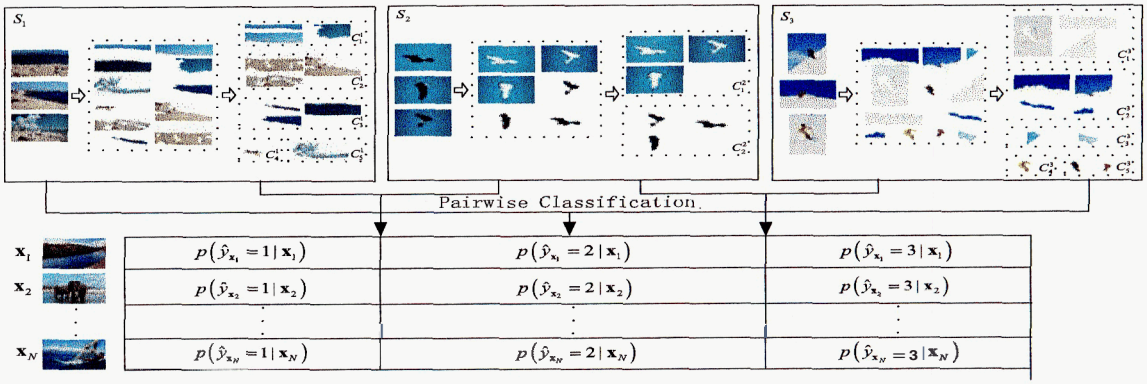


Figure 1. The work flow of the feature-to-concept mapping process.

adaptively determines the token number n_i by the ***Minimum Message Length (MML)*** criterion, and uses the *Expectation Maximization (EM)* algorithm to iteratively calculate the following parameters: the token priors $p(C_k^i)$, the mean vectors $\overline{\mu}_k^i$ and the co-variance matrixes Σ_k^i, $k = 1, \ldots, n_i$.

Given an image $\mathbf{x} \in X$ and its partitions $r_j(\mathbf{x}) \in \mathbf{x}$, we assume that the probability of $\mathbf{x}$ belonging to C_k^i is determined by its region partitions. Thus we have the conditional independence relationship $p(C_k^i|\mathbf{x}, r_j(\mathbf{x})) = p(C_k^i|\mathbf{x})$. Let $p(C_k^i|\mathbf{x})$ be the probability of $\mathbf{x}$ belonging to C_k^i, we have:

$$p(C_k^i|\mathbf{x}) = \sum_{r_j(\mathbf{x}) \in \mathbf{x}} p(r_j(\mathbf{x})|\mathbf{x}) p(C_k^i|r_j(\mathbf{x})) \tag{2}$$

where

$$p(C_k^i|r_j(\mathbf{x})) = \frac{p(C_k^i) p(r_j(\mathbf{x})|C_k^i)}{\sum_{g=1}^{n_i} P(C_g^i) p(r_j(\mathbf{x})|C_g^i)} \tag{3}$$

$p(r_j(\mathbf{x})|\mathbf{x})$ is the region saliency $v_j(\mathbf{x})$, $p(r_j(\mathbf{x})|C_k^i)$ and $p(C_k^n)$ are obtained by previous grouping process.

$P(C_k^i|\mathbf{x})$ can be interpreted as the appropriate degree of using C_k^i to represent image $\mathbf{x}$. Thus defining a vector $\mathbf{m}^i(\mathbf{x}) = [p(C_1^i|\mathbf{x}), \ldots, p(C_{n_i}^i|\mathbf{x})]^T$, $\mathbf{m}^i(\mathbf{x})$ can be treated as a set of real-valued features for $\mathbf{x}$ in the feature space spanned by $C_1^i, \ldots, C_{n_i}^i$, with each entry denoting how well each C_k^i interprets the image. The feature representation process can be intuitively explained as follows: the region of $\mathbf{x}$, $r_j(\mathbf{x})$, is represented in the feature space spanned by $C_1^i, \ldots, C_{n_i}^i$ as $\mathbf{G}^i(r_j(\mathbf{x})) = [p(C_1^i|r_j(\mathbf{x})), \ldots, p(C_{n_i}^i|r_j(\mathbf{x}))]^T$. Then $\mathbf{x}$ is represented as $\mathbf{m}^i(\mathbf{x}) = \sum_{r_j(\mathbf{x}) \in \mathbf{x}} p(r_j(\mathbf{x})|\mathbf{x}) \mathbf{G}^i(r_j(\mathbf{x}))$.

2.3. Generating posterior-based features

Given three semantic concepts S_i, S_j and S_k, the way that S_i differs from S_j is not the same in the way that S_i differs from S_k. Pairwise classification can be used to reasonably determine the optimal way to discriminate every pair of them. After pairwise classification, a posterior-based feature representation can be established for images.

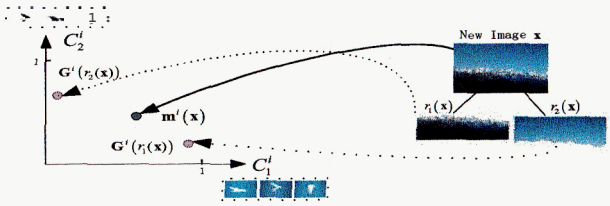


Figure 2. An example of representing a new image $\mathbf{x}$ in the feature space spanned by the two low-level tokens t_i, t_j which are specifically extracted for semantic concept c_i. $\mathbf{x}$ contains 2 regions $r_1(\mathbf{x})$ and $r_2(\mathbf{x})$. Image $\mathbf{x}$ is then represented as a real-valued feature vector $\mathbf{m}^i(\mathbf{x}) = \sum_{j=1,2} p(r_j(\mathbf{x})|\mathbf{x}) \mathbf{G}^i(r_j(\mathbf{x}))$.

To discriminate M semantic concepts, totally $M(M-1)/2$ painvise classifiers are generated. Each classifier is learned as follows. For two semantic concepts S_j and S_k, we can put images in a uniform feature space spanned by their low-level tokens. Given an image $\mathbf{x}$, we combine its feature vectors $\mathbf{m}^j(\mathbf{x})$ and $\mathbf{m}^k(\mathbf{x})$ together as $\tilde{\mathbf{m}}^{jk}(\mathbf{x})$:

$$\tilde{\mathbf{m}}^{jk}(\mathbf{x}) = \begin{bmatrix} \mathbf{m}^j(\mathbf{x}) \\ \mathbf{m}^k(\mathbf{x}) \end{bmatrix}^T \tag{4}$$

Put all images in $\mathcal{X}^j$ and $\mathcal{X}^k$ together as $\mathcal{X}^{jk}$, and represent each image $\mathbf{x} \in \mathcal{X}^{jk}$ by the new $\tilde{\mathbf{m}}^{jk}(\mathbf{x})$. Then a classifier can be constructed based on $\mathcal{X}^{jk}$ to separate S_j and S_k. Let $\hat{y}_\mathbf{x}$ be the estimated class label of $\mathbf{x}$, and let

$$\xi^{jk}(\mathbf{x}) = p(\hat{y}_\mathbf{x} = j | \mathbf{x} \in \mathcal{X}^j \text{ or } \mathbf{x} \in \mathcal{X}^k) \tag{5}$$

be the class probability estimated by the classifier between S_j and S_k. We have $p(\hat{y}_\mathbf{x} = k | \mathbf{x} \in \mathcal{X}^j \text{ or } \mathbf{x} \in \mathcal{X}^k) = 1 - \xi_i^{jk}$. Then the class probabilities estimated by the $M(M-1)/2$ painvise classifiers can be combined to give an ensemble class probability estimation $p(\hat{y}_\mathbf{x} = j|\mathbf{x})$ for any image $\mathbf{x}$ in the database. In this paper, we use the probability estimation method proposed in [9] to combine the hypotheses:

$$p(\hat{y}_\mathbf{x} = j|\mathbf{x}) = \frac{1}{Z} \cdot \frac{1}{\sum_{k:k \neq j} \frac{1}{\xi^{jk}(\mathbf{x})} - (M-2)} \tag{6}$$

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

where Z normalizes $\sum_{j=1}^{M} p(\hat{y}_{\mathbf{x}} = j|\mathrm{x})$ to 1. Vector $[p(\hat{y}_{\mathbf{x}}=1|\mathbf{x}), \ldots, p(\hat{y}_{\mathbf{x}}=M|\mathbf{x})]^T$ can be treated as a posterior-based feature vector of image x represented by M semantic concepts. This feature is high-level feature because it is better related to high-level semantic concepts than the original low-level features. This posterior-based high-level feature is generated based on low-level tokens, which **is** extracted from region-based low-level features. Thus we call the whole process the ***mapping*** of ***low-level visualfeatures to high-level semantic concepts.***

3. Supervisory Information from LRF

LRF is an effective mechanism to accumulate high-level semantic information from retrievals by previous users [4,5]. In *[5]* a ***multi-layer semantic representation (MSR)*** **is** proposed to describe the real-world semantics of images, and an automatic algorithm is implemented to extract the MSR through LRF process. The MSR records direct multi-correlations among images, and integrates these records to extract hidden semantic concepts from the database. The extracted concepts are distributed in multiple semantic layers, with one semantic layer corresponding to one kind of hard partitions of the semantic space. We do not verbosely describe the detailed process of generating the MSR. The information contained in the **MSR** is what we need.

Assume that the system extracts M semantic concepts $S_1, \ldots, S_M$ in the MSR. In [5], each concept S_i is represented by a quadruple $(\mathcal{X}^i, \overline{\mathcal{X}}^i, \mathcal{H}^i, \overline{\mathcal{H}}^i)$. $\mathcal{X}^i = \{\mathbf{x}_1^{i+}, \ldots, \mathbf{x}_{n_i}^{i+}\}$ is the labeled "relevant" image set for S_i; $\mathcal{H}^i = \{h_1^{i+}, \ldots, h_{n_i}^{i+}\}$, where h_j^{i+} is the the counting number of image $\mathbf{x}_j^{i+}$ being labeled to be in $\mathcal{X}^i$. Similarly $\overline{\mathcal{X}}^i = \{\mathbf{x}_1^{i-}, \ldots, \mathbf{x}_{m_i}^{i-}\}$ and $\overline{\mathcal{H}}^i = \{h_1^{i-}, \ldots, h_{m_i}^{i-}\}$ are the labeled "irrelevant" image set and the counting number of each image in $\overline{\mathcal{X}}^i$ being labeled to be in $\overline{\mathcal{X}}^i$ respectively. The learned concept information $S_1, \ldots, S_M$ and "relevant" sets $\mathcal{X}^i$, $i = 1, \ldots, M$ can be used as the supervisory information to extract low-level tokens for the learned concepts, and to generate posterior-based features through pairwise classification.

4. The Integrated Image Retrieval

In this section, we implemented an integrated system, which seamlessly combines SRF with LRF in our system.

4.1. Retrieval by SRF and LRF

Assume that during a retrieval process, in feedback round t, images in sets $\mathcal{R}^t$ and $\mathcal{IR}^t$ are labeled to be "relevant" and "irrelevant" respectively by the user. For an image x in the database, let $E(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ denote the probability that it fits the current query concept. $E(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ is given by:

$$E(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t) = \lambda p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t) + (1-\lambda) p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t) \quad (7)$$

where $p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ is the probability estimated by offline feature-to-concept mapping, and $p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ is the probability predicted by cumulate long-term records.

In practical retrieval, $p(S_i)$ and $p(\mathrm{x})$ follow uniform distribution. We write $p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ and $p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ as:

$$p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t) = \frac{1}{Z_s} \sum_{i=1}^{M} p_s(S_i|\mathcal{R}^t, \mathcal{IR}^t) \cdot p_s(S_i|\mathbf{x}) \quad (8)$$

$$p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t) = \frac{1}{Z_l} \sum_{i=1}^{M} p_l(S_i|\mathcal{R}^t, \mathcal{IR}^t) \cdot p_l(S_i|\mathbf{x}) \quad (9)$$

where Z_s and Z_l respectively are normalization factors for $p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ and $p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$. $p_s(S_i|\mathcal{R}^t, \mathcal{IR}^t)$ is the probability that the current query concept fits the concept S_i, estimated by offline feature-to-concept mapping. $p_l(S_i|\mathcal{R}^t, \mathcal{IR}^t)$ is the probability estimated by long-term memory. They can be given by:

$$p_s(S_i|\mathcal{R}^t, \mathcal{IR}^t) = \frac{1}{Z'_s} \prod_{\mathbf{x}_j \in \mathcal{R}^t} p_s(S_i|\mathbf{x}_j) \prod_{\mathbf{x}_j \in \mathcal{IR}^t} [1 - p_s(S_i|\mathbf{x}_j)] \quad (10)$$

$$p_l(S_i|\mathcal{R}^t, \mathcal{IR}^t) = \frac{1}{Z'_l} \prod_{\mathbf{x}_j \in \mathcal{R}^t} p_l(S_i|\mathbf{x}_j) \prod_{\mathbf{x}_j \in \mathcal{IR}^t} [1 - p_l(S_i|\mathbf{x}_j)] \quad (11)$$

where Z'_s and Z'_l respectively are normalization factors for $p_s(S_i|\mathcal{R}^t, \mathcal{IR}^t)$ and $p_l(S_i|\mathcal{R}^t, \mathcal{IR}^t)$. In fact for any image x, $p_s(S_i|\mathrm{x}) = p_s(\hat{y}_{\mathbf{x}} = i|\mathrm{x})$, $p_l(S_i|\mathrm{x}) = p_l(\hat{y}_{\mathbf{x}} = i|\mathrm{x})$. The offline feature-to-concept mapping process and the long-term records give $p_s(\hat{y}_{\mathbf{x}} = i|\mathbf{x})$ and $p_l(\hat{y}_{\mathbf{x}} = i|\mathrm{x})$, respectively, as:

- Offline feature-to-concept mapping:

$p_s(\hat{y}_{\mathbf{x}} = i|\mathrm{x})$ is given by the ensemble class probability described in Eqn.(6):

$$p_s(\hat{y}_{\mathbf{x}} = i|\mathrm{x}) = \frac{1}{Z} \cdot \frac{1}{\sum_{k:k\neq i} \frac{1}{\xi^{ik}(\mathbf{x})} - (M-2)}$$

- Long-term memory:

$p_l(\hat{y}_{\mathbf{x}} = i|\mathrm{x})$ is given by:

$$p_l(\hat{y}_{\mathbf{x}} = i|\mathbf{x}) = h^{i+} / \sum_{k=1}^{M} h^{k+} \quad (12)$$

where h^{i+} is the count of the image x being labeled to be in the "relevant" set $\mathcal{X}^i$ for concept S_i.

With Eqns.(6,10), $p_s(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ can be obtained by Eqn.(8). With Eqns.(12,11), $p_l(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ can be obtained by Eqn.(9). Then the integrated $E(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$ can be calculated. During the feedback round t, the images in the database are ranked based on $E(\mathbf{x}|\mathcal{R}^t, \mathcal{IR}^t)$, and those with larger $E(\mathrm{x}|\mathcal{R}^t, \mathcal{IR}^t)$ are returned as the retrieval result.

Eqn.(7) is an integrated retrieval strategy, which combines the posterior-based features from the feature-to-concept mapping process and the semantic information from the accumulated LRF records to improve the retrieval performance. λ is a parameter to adjust the relative importance of these two **kinds** of information.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

4.2. System implementation

By now we give the entire scheme of our RBIR system in Fig.3. In the first round of retrieval, there is only a query image $\mathbf{x}_q$ as the positive sample, the system retrieves images by calculating their similarities to the query image and then ranking the similarities in descending order. The similarity is given by the ***Unified Feature Matching*** (UFM) measurement described in [2]. Feature-to-concept mapping is carried out when every 100 retrievals are taken by the users.

In here, the *Support* ***Vector Machine*** (***SVM***) is adopted as the pairwise classifier mentioned in Sec.2.3 The output of the **SVM** classifier is the distance $D^{jk}(\mathbf{x})$ of image $\mathbf{x}$ from the classification boundary. $D^{jk}(\mathbf{x})$ can be converted to probability $\xi^{jk}(\mathbf{x})$ in Eqn.(5) as:

$$\xi^{jk}(\mathbf{x}) = \frac{1}{1+\exp\{-D^{jk}(\mathbf{x})\}} \tag{13}$$

Initialization: User provides a query image $\mathbf{x}_q$ **to start** retrieval. The first round of retrieval is conducted by ranking images according to their UFM similarity to $\mathbf{x}_q$.

Iteration: for $=1$

1. Calculate $_s(\mathbf{x}|\mathcal{R}^t\ \mathcal{IR}^t)$ by Eqn.(8) with Eqns.(6,10);
2. Calculate $_I(\mathbf{x}|\mathcal{R}^t\ \mathcal{IR}^t)$ by Eqn.(9) with Eqns.(12,11);
3. Calculate $(\mathbf{x}|\mathcal{R}^t\ \mathcal{IR}^t)$ by Eqn.(7);
4. Rank images by $(\mathbf{x}|\mathcal{R}^t\ \mathcal{IR}^t)$ in descending order;
5. If user is satisfied, stop iteration; otherwise, user labels the top images which has not been labeled during this query session as feedback information.

Terminal: Long-term MSR learning by algorithm in [5].

Figure 3. Retrieval algorithm in our system.

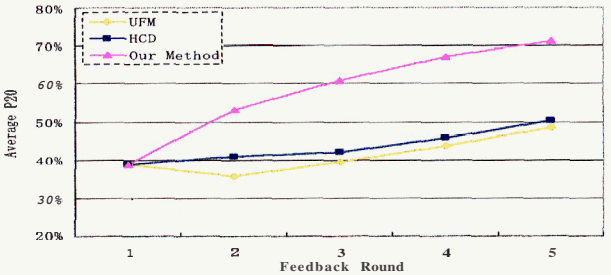
5. Experimental Results

The image database used in our experiments has 10,000 real-world images from Corel CDs, in which there are 100 semantic categories, 100 images per category. After segmentation there are 73,083 regions for the entire database. The color features used are the first three color moments in LUV space (9 dimensions), and the color histogram in HSV space (**64** dimensions); the texture features used are the coarseness vector (10 dimensions), directionality (8 dimensions) and tree-structured Wavelet transform texture (104 dimensions). We perform 2000 rounds of simulated retrieval based on the ground-truth concepts to accumulate the semantic concepts and the "relevant" training set for each concept through the LRF learning. **In** the experiments, the performance measurement used is the top-k precision P_k (the percentage ***of*** "relevant" images in the returned k images). The user labels 10 images during each feedback round.

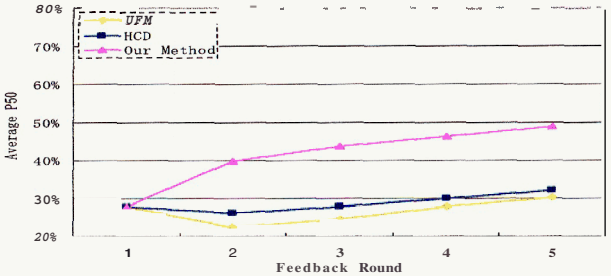
5.1. Comparison with the state-of-the-arts

To clearly evaluate the performance of the offline learning process which maps low-level features to high-level semantic concepts through extracting low-level tokens, we set $\lambda = 1$ in the algorithm described in Fig.3 (the long-term records are not used), and compare our method with two other RBIR approaches: the classical ***Unified Feature Matching*** (***UFM***) method **[2]** (a typical method to retrieve images based on directly calculating image-to-image similarity); and the unsupervised hidden concept discovery (*HCD*) method [15] (a typical method to retrieve images based on extracting new features in the form of offline unsupervised learning). In this experiment, 1000 query images are randomly selected for 1000 retrievals, and the average P_k **is** calculated. For a fair comparison, all the methods use the same low-level features, the same segmentation results, and the same query images. Moreover for all methods, the first round of retrieval is carried out by ranking images according to their UFM similarity to the query image.

Fig.4 gives the average P_{20} and P_{50} of our method, UFM and HCD. The dimensionality of the extracted new features for HCD [15] is set to 800. From the figure, we can see that both HCD and our method are better than the UFM approach. This indicates that the method to extract new features for retrieval is more effective than the method directly using the original region-based low-level features. Furthermore, our method outperforms HCD significantly and consistently from the second feedback round. For example, the precision improvement in round 5 attains 41%. This shows that the supervised feature extraction and feature selection can better map low-level features to high-level semantic concepts than the unsupervised feature extraction approach.



(a) Average 20



@)Average 60

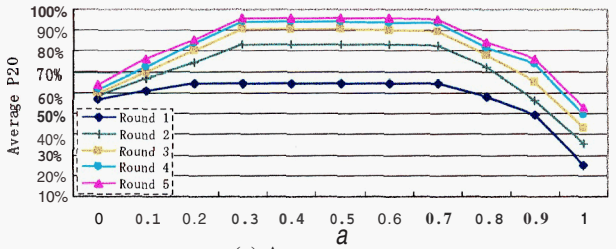
Figure 4. Comparison **of** our method and UFM and HCD.

5.2. Evaluation of the integrated retrieval

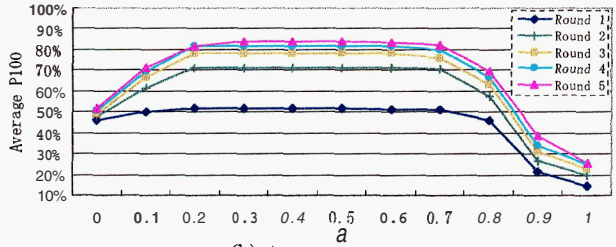
In this experiment, we evaluate the performance of the integrated retrieval strategy which incorporates both infor-

mation from feature-to-concept mapping and that from cumulate LRF records. The long-term experience usually yields good results when query images have already been recorded, and does not work well when query images have not. To make the evaluation more reasonable, 2000 query sessions are conducted to calculate the average precision as follows. 1000 images are randomly selected from the images which have already been recorded in cumulate memory, 1000 images are randomly selected from the remaining images which are not recorded in previous memory.

Fig.5 gives the average P_{20} and P_{100} of our proposal with λ varying from 0 to 1. The figure shows that both the long-term records and information from the feature-to-concept mapping process can improve the retrieval performance. When $\lambda < 0.3$ or $\lambda > 0.7$, the system almost uses only the offline learned information from the mapping process or only the LRF learned records for retrieval, and the performance deteriorates rapidly. When $0.3 < \lambda < 0.7$, the retrieval performance is stable, which indicates that the system is not sensitive to λ when it takes value in a large range.



(a) Average $_{20}$



(b) Average $_{100}$

Figure 5. Retrieval accuracy with varying from 0 to 1.

6. Conclusions

We propose a novel offline supervised learning method to map low-level features to high-level concepts through extracting low-level tokens in RBIR context. According to the assumption that regions from images in the same semantic concept are generatively formed by low-level tokens, different low-level tokens are extracted for different concepts respectively. Representing images in the new feature space spanned by low-level tokens, pairwise classifiers are constructed to discriminate each pair of concepts and generate posteriors for images, which are treated as high-level features connecting images to concepts. LRF is incorporated to provide supervisory information. An integrated algorithm is implemented to combine the information from feature-to-concept mapping and that from LRF memory for retrieval. Experimental results demonstrate the effectiveness of our proposed approach.

References

[1] C. Carson, *et al.*, "Blobworld: A system for region-based image indexing and retrieval," ***Proc. Int. Conf. on Visual Information System,*** pp.509-516, 1999.

[2] Y.X. Chen, J.Z. Wang, "A region-based fuzzy feature matching approach for content-based image retrieval," ***IEEE Trans. on PAMI,*** 24(9):1252-1267, 2002.

[3] A.L. Dong, B. Bhanu, "A new semi-supervised EM algorithm for image retrieval," ***Proc. IEEE Int. Con$ on CVPR,*** vol.2, pp.662-667, 2003.

[4] X.F. He *et al.*, "Learning a semantic space from user's relevance feedback for image retrieval", ***IEEE Trans. on CSVT,*** 13(1):39-48, 2003.

[5] W. Jiang, G.H. Er, Q.H. Dai, "Multi-layer semantic representation learning for image retrieval", ***IEEE Int. Conf. on Image Processing,*** TP P6.5, Singapore, Oct. 2004.

[6] F. Jing, *et al.*, "An effective and efficient region-based image retrieval framework," ***IEEE Trans. on Image Processing,*** 13(5):699-709, 2004.

[7] M.H.C. Law, *et al.*, "Simultaneous feature selection and clustering using mixture models," ***IEEE Trans. on PAMI,*** 26(9): 1154-1166, 2004.

[8] Y.F. Ma, H.J. Zhang, "Contrast-based image attention analysis by using fuzzy growing", ***Proc. ACM Int. Conf. on Multimedia,*** pp.374-381, 2003.

[9] D. Price, *et al.*, "Pairwise neural network classifiers with probabilistic outputs," G. Tesauro, D. Touretzky and T. Leen editors, ***Neural Information Processing Systems,*** vol.7, pp.1109-1116, The MIT Press, 1995.

[10] Y. Rui, *et al.*, "Relevance feedback: A powerful tool in interactive content-based image retrieval", ***IEEE Trans. on CSVT, Special Issue on Segmentation, Description and Retrieval of Video Content,*** 8(5):644-655, 1998.

[11] A.W.M. Smeulders, *et al.*, "Content-based image retrieval at the end of the ealy years", ***IEEE Trans. on PAMI,*** 22(12):1349-1380, 2002.

[12] K. Tieu, P. Viola, "Boost image retrieval", ***Proc. IEEE Int. Conf. on CVPR,*** vol.1, pp.228-235, 2000.

[13] J.Z. Wang, *et al.*, "Simplicity: Semantic-sensitive integrated matching for picture libraries," ***IEEE Trans. on PAMI,*** 23(9):947-963, 2001.

[14] Y. Wu, *et al.*, "Discriminant EM algorithm with application to image retrieval", ***Proc. IEEE Int. Conf. on CVPR,*** vol.1, pp.222-227, 2000.

[15] R.F. Zhang, Z.F. Zhang, "Hidden semantic concept discovery in region based image retrieval," ***Proc. IEEE Int. Con. CVPR,*** ~01.2pp.996-1001, 2004.

[16] X. Zheng, X.Y. Lin, "Automatic determination of intrinsic cluster number sequence in spectral clustering using random walk on graph," ***IEEE Int. Conf. on Image Processing,*** WP P8.2, Singapore, Oct. 2004.