Visual Keyword Based Image Retrieval

Yeong-Yuh Xu[1] and Hsin-Chia Fu[2]

**[1] Department of Computer Science and Information Engineering, HungKuang University
Taichung, Taiwan, R.O.C.**

**[2] College of Engineering, Huaqiao University,
Quanzhou, China**

Abstract

This paper proposes the *visual keywords* to represent the visual appearance, such as color and texture, of regions in images without precisely image segmentation. Based on the proposed *visual keywords*, a multi-modal image query and retrieval (MIQR) approach is presented to retrieve desired images from image databases by using textual annotations associated with images and/or the proposed *visual keyword*. In addition, the Generalized Probabilistic Decision-Based Neural Networks is adopted to model the *visual keywords*. The experiments were performed on a subset of the COREL image gallery and the images gathered from Internet. A comparison with current leading approaches is made. The experimental results show that the MIQR approach can retrieve relevant images closely associated with users' query regions or objects, and has the capability of searching and retrieving relevant images from a large collection of images.

Keywords: *Content-based image retrieval, Visual keywords, Multi-modal image query and retrieval, Generalized Probabilistic Decision-based Neural Networks, GPDNN.*

1. Introduction

The ongoing proliferation of digital media available over Internet, in enterprises, or at home, has spawned great demand for the navigation and searching of desired visual contents in a large database [1]. Traditional database techniques have been adequate for many applications involving text records which could be ordered, indexed and search, for matching patterns in a straightforward manner. However, text-based search techniques are in principle in adequate for non-textual unstructured content, such as image data. Although it works fairly well for retrieving images with text annotations such as named entities, e.g., specific people, objects, or places, the text-based approach does not work well for generic topics related to general settings of objects in images since the text annotation rarely describes the background setting or the visual appearance, such as color, texture, and shape of the objects. Apparently, traditional text-based search techniques without considering the visual appearance are not sufficient for effective image retrieval. Hence, a seamlessly integrated paradigm of text-based and visual content-based query is needed for image retrieval on general propose multimedia databases.

In recent years, there has been a dramatic proliferation of research concerned with the image retrieval. These research studies can be roughly categorized into two categories:

1.1 Text-Based Methods

The texts surrounding multimedia objects are analyzed, and those appeared to be relevant are extracted. Shen, Ooi, and Tan [2] explored the context of web pages as potential annotations for the images in the same pages. Srihari et al. [3] proposed extracting named entities from the surrounding text to index images. The major constraint of text-based methods is that it requires the presence of high quality textual information in the surrounding of the multimedia objects. In many situations, this requirement may not be satisfied.

1.2 Content-Based Methods

In the content-based image retrieval (CBIR), images are indexed and retrieved by their visual features, such as object shape, position, color, texture, etc. [4, 5, 6, 7, 8, 9, 10]. Some developed image retrieval systems [11, 12] basically applied global features, such as color histograms or transform coefficient histograms. However, using global features for image indexing, a query system may ignore some significant local details of an image, so as to retrieve undesired images. Instead of using global features of an image, some query systems [13, 14, 15] adopted local features to represent or to index an image. In these systems, the local features are obtained from regions or subimages which are segmented or sketched from an image first, and then various visual features of these regions are extracted. In general, the query and retrieving precision of these systems are usually better than the global feature based systems, but their performance

depend heavily on the segmenting or sketching accuracy of the regions.

More methods focus on extracting semantic information directly from the content of multimedia objects [16, 17, 18, 19]. Until very recently, segmenting an image into semantic meaningful objects is still a difficult task in image processing [4, 5, 6]. Instead of emphasizing on the precise region segmentation, Wang, Li, and Wiederhold [7] proposed SIMPLIcity, an image retrieval system capturing semantics using the robust Integrated Region Matching metric. The semantics are used to classify images into two broad categories, and then used to support semantics-sensitive image retrievals. Recently, Goh, Li, and Chang [20] proposed a confidence-based dynamic ensemble (CDE) to use local and global perceptual features to annotate images. CDE can make dynamic adjustments to accommodate new semantics, to assist the discovery of useful low-level features, and to improve class-prediction accuracy. However, if images have neither dominant regions nor common visual features, these methods may probably fail to conclude an acceptable result, i.e., a semantic meaning.

In this paper, the *visual keyword* is proposed to represent an image region without precisely image segmentation, and the Generalized Probabilistic Decision-Based Neural Networks (GPDNN) is adopted to model the proposed *visual keyword*. Furthermore, a multi-modal image query and retrieval (MIQR) approach is presented to make the query result acceptable by seamlessly integrating paradigm of text-based and *visual keyword*-based query and retrieval.

The reminder of this paper is organized as follows. In Section 2, the proposed *visual keyword* is presented. Section 3 contains a description of the proposed multi-modal image query and retrieval (MIQR) approach. Finally, experimental results are provided in Section 4, while conclusions are reported in Section 5.

2. Visual keywords

Since the text-based query is not mature or effective on non-textual unstructured content, the *visual keywords* are proposed to describe the contents of images. As similarity to a document made up of the words, an image is composed of a set of homogeneous regions that collectively cover the entire image. The pixels in the same homogeneous region are similar with respect to some characteristic or computed property, such as color, texture, etc. In order to represent the homogeneous regions that may (or may not) be segmented imprecisely, the proposed *visual keywords* roughly illustrate the characteristics of the regions. The details of the *visual keywords* are described as follows.

2.1 Generation

Suppose that we want to illustrate the characteristics of a homogeneous region A_i in an image I. A *visual keyword* ω_i is defined as a mixture of Gaussian distributions to formulate the color, texture, and spatial features of the region A_i. What needs to be emphasized is that approximating A_i by the mixture of Gaussian distributions has two beneficial effects: (1) avoiding the complicate and difficult image segmentation process and (2) giving more flexibility in matching with the other similar regions. Following will first introduce how we determine the homogeneous regions in the images, and then describe how we generate the *visual keywords* from the homogeneous regions.

As noted previously, the homogeneous region contains pixels of similar color and texture. Since the Color Measurement Committee (the CMC) distance [21] in the CIELab color space is one of most accurate metrics according to human perception, we employ the CMC distance to measure the color difference between pixels. Because the Gabor representation is optimal [22] in the sense of minimizing the uncertainty in the space and the frequency domain, the texture features of pixels are extracted from the image of multiple scales and orientations by using the Gabor wavelet decomposition [23]. Based on these visual features, an expansion process is used to make the homogeneous region ``grow" gradually from a given reference pixel. The expansion process continues to check the neighborhood of the expanded region until the color and texture features of the neighborhood are far different from the reference pixel. Finally, we obtain a homogeneous region A_i, the surrounding area of the reference pixel with closer color and similar texture features. After the homogeneous region A_i is determined, a *visual keyword* ω_i to represent A_i is generated by the following two stages: (1) the spatial modeling and (2) the color and texture modeling.

2.1.1 The spatial modeling

Since the homogeneous region is an arbitrary shaped region, it can be approximated by the union of several elliptic regions. Given a homogeneous region A_i containing a set of pixels $X = \{\mathbf{x}(t) : t = 1, 2, \ldots, N\}$, we assume that the union of elliptic regions to approximate A_i is a mixture of Gaussian distributions

$$p_s(\mathbf{x}(t)|\ \omega_i)=\sum_{r=1}^{R}P_s(\theta_{sr}|\ \omega_i)p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr}),$$

where θ_{sr} represents the r th cluster, and $P_s(\theta_{sr}|\ \omega_i)$ denotes the prior probability of the cluster r. By definition, $\sum_{r=1}^{R}P_s(\theta_{sr}|\ \omega_i)=1$, where R is the number of clusters in $p_s(\mathbf{x}(t)|\ \omega_i)$. Define $p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})$ to be one of the Gaussian distributions which comprise $p_s(\mathbf{x}(t)|\ \omega_i)$. Since the homogeneous region is a two-dimensional region, the cluster r is a 2 D Gaussian distribution:

$$p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})=\frac{1}{2\pi|\Sigma_{sr}|^{1/2}}\times\exp\left(-\frac{1}{2}(\mathbf{x}(t)-\boldsymbol{\mu}_{sr})^T\Sigma_{sr}^{-1}(\mathbf{x}(t)-\boldsymbol{\mu}_{sr})\right), \quad (1)$$

where $\boldsymbol{\mu}_{sr}=[\mu_{sr1},\mu_{sr2}]^T$ is the mean vector, and Σ_{sr} is the covariance matrix. The EM algorithm [24] is applied to adjust the parameters of Eq. (1). In each EM iteration, there are two steps: Estimation (E) step and Maximization (M) step. The M step maximizes a likelihood function which is further refined in each iteration by the E step. The goal of the EM learning is to maximize the log likelihood of the pixels in A_i, $E=\sum_{t=1}^{N}\log(p_s(\mathbf{x}(t)|\ \omega_i))$. The following update equations are used to adjust the parameters of 2 D mixture Gaussian distribution: at each iteration j,

$$\mu_{sr}^{(j+1)}=\frac{\sum_{t=1}^{N}h_{sr}^{(j)}(t)\mathbf{x}(t)}{\sum_{t=1}^{N}h_{sr}^{(j)}(t)},$$

$$\Sigma_{sr}^{(j+1)}=\frac{\sum_{t=1}^{N}h_{sr}^{(j)}(t)(\mathbf{x}(t)-\mu_{sr}^{(j)})(\mathbf{x}(t)-\mu_{sr}^{(j)})^T}{\sum_{t=1}^{N}h_{sr}^{(j)}(t)},$$

$$P_s^{(j+1)}(\theta_{sr}|\ \omega_i)=\frac{1}{N}\sum_{t=1}^{N}h_{sr}^{(j)}(t),$$

where

$$h_{sr}^{(j)}(t)=\frac{P_s^{(j)}(\theta_{sr}|\ \omega_i)p_s^{(j)}(\mathbf{x}(t)|\ \omega_i,\theta_{sr})}{\sum_{r=1}^{R}P_s^{(j)}(\theta_{sr}|\ \omega_i)p_s^{(j)}(\mathbf{x}(t)|\ \omega_i,\theta_{sr})}.$$

When the EM iteration converges, it should ideally obtain maximum likelihood estimation of the data distribution. Figure 1 shows an example of the spatial modeling of a sail boat. The original image with a reference pixel (the black dot) is depicted in Figure 1(a), and the corresponding homogenous region and its spatial model is depicted in Figure 1(b) and (c), respectively.

2.1.2 The color and texture modeling

As mentioned above, the spatial features of the homogeneous region A_i in the image I are approximated by the union of R elliptic regions. After the spatial modeling, our approach models color and texture in each elliptic region. Suppose that an elliptic region a_r is approximated by a Gaussian distribution $p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})$. Then, the color features of a_r is modeled by a Gaussian distribution:

$$p_c(\mathbf{c}_{\mathbf{x}(t)}|\ \omega_i,\theta_{cr})=\frac{1}{(2\pi)^{D_c/2}|\Sigma_{cr}|^{1/2}}\times\exp\left(-\frac{1}{2}(\mathbf{c}_{\mathbf{x}(t)}-\boldsymbol{\mu}_{cr})^T\Sigma_{cr}^{-1}(\mathbf{c}_{\mathbf{x}(t)}-\boldsymbol{\mu}_{cr})\right),$$

where $\mathbf{c}_{\mathbf{x}(t)}$ is a D_c-dimensional color feature vector at a pixel $\mathbf{x}(t)$, and the mean vector $\boldsymbol{\mu}_{cr}$ and the covariance matrix Σ_{cr} are calculated as follows:

$$\boldsymbol{\mu}_{cr}=\frac{1}{N_{cr}}\sum_{\mathbf{x}(t)\in I}p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})\mathbf{c}_{\mathbf{x}(t)}$$

$$\Sigma_{cr}=\frac{1}{N_{cr}}\sum_{\mathbf{x}(t)\in I}p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})(\mathbf{c}_{\mathbf{x}(t)}-\boldsymbol{\mu}_{cr})(\mathbf{c}_{\mathbf{x}(t)}-\boldsymbol{\mu}_{cr})^T,$$

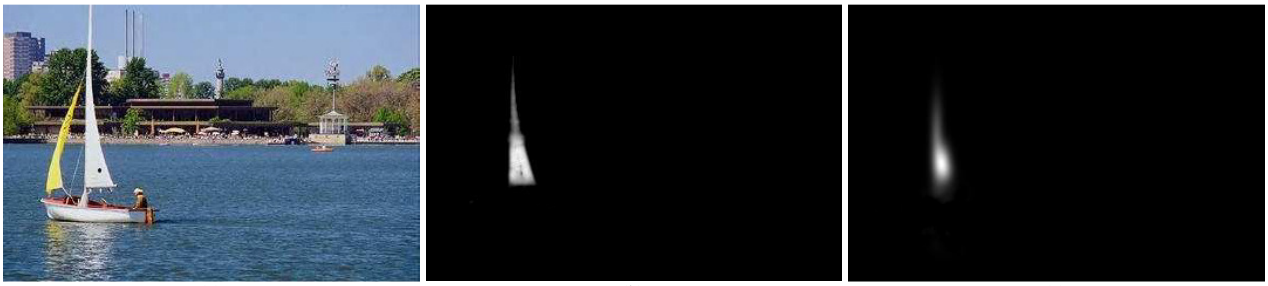
where $N_{cr}=\sum_{\mathbf{x}(t)\in I}p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})$. Since the texture features of elliptic regions are modeled in the same way as color features, one can obtain the notations and formulas for texture modeling by just replacing the term ``color'' by ``texture'' and the subscript c by t in the above description. After the spatial, color, and texture modeling, the joint Gaussian density function is obtained as

$$p(\mathbf{z}(t)|\ \omega_i,\theta_r)=p_s(\mathbf{x}(t)|\ \omega_i,\theta_{sr})p_c(\mathbf{c}_{\mathbf{x}(t)}|\ \omega_i,\theta_{cr})p_t(\mathbf{t}_{\mathbf{x}(t)}|\ \omega_i,\theta_{tr}),$$

where $\mathbf{z}(t)=[\mathbf{x}(t)^T,\mathbf{c}_{\mathbf{x}(t)}^T,\mathbf{t}_{\mathbf{x}(t)}^T]^T$, and $\theta_r=[\theta_{sr},\theta_{cr},\theta_{tr}]$.

Finally, the *visual keyword* ω_i formulates the spatial, color, and texture features of the homogenous region A_i by a mixture of Gaussian distributions

$$p(\mathbf{z}(t)|\ \omega_i)=\sum_{r=1}^{R}P_s(\theta_{sr}|\ \omega_i)p(\mathbf{z}(t)|\ \omega_i,\theta_r),$$



(a) (b) (c)

Fig. 1 The spatial modeling of a sail region (a) The original image with a reference pixel showing as a black dot on a sail (b) The corresponding homogeneous region (c) The 2D mixture Gaussian distributions to approximate the sail region

2.1 Visual keywords modeling

Due to the distributional forms of the {\it visual keywords}, the Generalized Probabilistic Decision-Based Neural Networks (GPDNN) [25] is adopted to model the {\it visual keywords}. The GPDNN is a generalized model from its predecessor, SPDNN [26]. The GPDNN contrasts with the SPDNN in respect to the input data type. The input data of the GPDNN are in the distributional forms instead of the numerical forms.

The schematic of a GPDNN is depicted in Figure 2. As similar to its predecessor SPDNN, GPDNN has a modular network structure, where one subnet is designated to model one particular class (*visual keyword*).

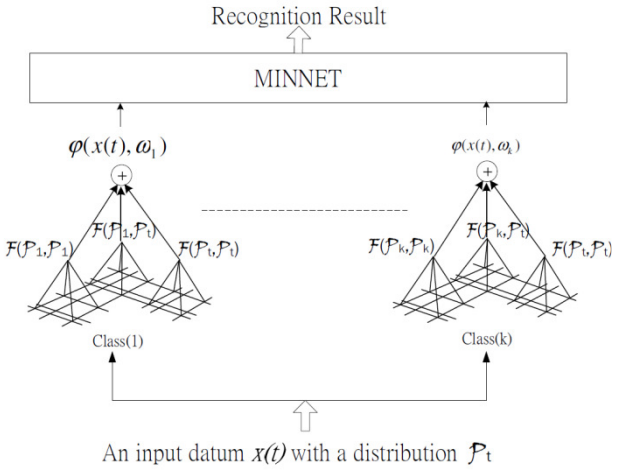


Fig. 2 The schematic diagram of a *k-class* GPDNN. The detail of a subnet is shown in Figure 3.

Assume that the modeled distribution for each class ω_i is $\mathcal{P}_i$, where $i \in \{1,2,\cdots,k\}$. For an input datum $\mathbf{x}(t)$ with a distribution $\mathcal{P}_t$, the discriminate function of GPDNN is defined as

$$\varphi(\mathcal{P}_i,\mathcal{P}_t) = \mathcal{F}(\mathcal{P}_i,\mathcal{P}_i) - 2\mathcal{F}(\mathcal{P}_i,\mathcal{P}_t) + \mathcal{F}(\mathcal{P}_t,\mathcal{P}_t),$$

where

$$\mathcal{F}(\mathcal{P}_i,\mathcal{P}_t) = \sum_{n=1}^{R_i}\sum_{m=1}^{R_t} P_n^i P_m^t \mathcal{G}(\theta_n^i,\theta_m^t),$$

and

$$\mathcal{G}(\theta_n^i,\theta_m^t) = \frac{\exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(\mu_{m(d)}^t-\mu_{n(d)}^i)^2}{(\sigma_{m(d)}^t)^2+(\sigma_{n(d)}^i)^2}\right\}}{\sqrt{\prod_{d=1}^{D}2\pi((\sigma_{m(d)}^t)^2+(\sigma_{n(d)}^i)^2)}}. \qquad (2)$$

R^D is a D-dimensional feature space; $\mathcal{P}_i = \sum_{n=1}^{R_i} P_n^i p(\mathbf{z}|\ \theta_n^i)$ and $\mathcal{P}_t = \sum_{m=1}^{R_t} P_m^t p(\mathbf{z}|\ \theta_m^t)$ are two mixture Gaussian distributions. Here, let * denote i or t, and R_* is the number of mixture components in $\mathcal{P}_*$, P_n^* is the prior probability of the n 'th component, $\theta_n^* = \{\mu_n^*, \Sigma_n^*\}$, and $p(\bullet|\ \theta_n^*)$ is a multivariate Gaussian with mean vector

$$\mu_n^* = [\mu_{n(1)}^*,\cdots,\mu_{n(D)}^*]^T,$$

and a covariance matrix

$$\Sigma_n^* = diag[(\sigma_{n(1)}^*)^2,\cdots,(\sigma_{n(D)}^*)^2].$$

Therefore, the discriminate function is implemented by a two-layer pyramid network. The bottom layer contains three structurally identical pyramid subnetworks, each of which computes the $\mathcal{F}(\mathcal{P}_i,\mathcal{P}_i)$, $\mathcal{F}(\mathcal{P}_i,\mathcal{P}_t)$, and $\mathcal{F}(\mathcal{P}_t,\mathcal{P}_t)$, respectively. Figure 3 depicts the internal architecture of the pyramid subnetwork corresponding to $\mathcal{F}(\mathcal{P}_i,\mathcal{P}_t)$. Suppose that the mixture Gaussian distributions $\mathcal{P}_i$ and $\mathcal{P}_t$ consist of R_i and R_t components, respectively. The subnetwork for $\mathcal{F}(\mathcal{P}_i,\mathcal{P}_t)$ contains R_i hidden nodes and a $R_i \times R_t$ input nodes, each of which is marked as $G_{n,m}$.

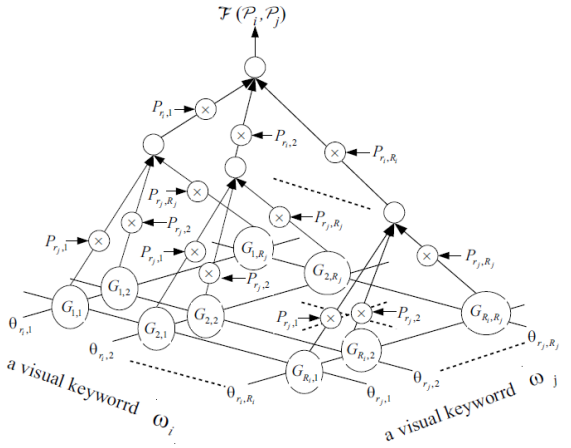


Fig. 3 The internal architecture of a model in Figure 2 for computing of $\mathcal{F}(\mathcal{P}_i, \mathcal{P}_i)$.

The GPDNN adopts the SPDNN learning scheme. While the input datum $\mathbf{x}(t)$ belonging to the class ω_i is misclassificated to the class w_j, the reinforced and antireforced learning rules are applied to the subnets of ω_i and w_j, respectively;

Reinforced Learning rule:

$$\mathbf{w}_i^{(m+1)} = \mathbf{w}_i^{(m)} + \eta \nabla \varphi(\mathbf{x}(t), \omega_i) \quad (3)$$

Antireinforced Learning rule:

$$\mathbf{w}_j^{(m+1)} = \mathbf{w}_j^{(m)} - \eta \nabla \varphi(\mathbf{x}(t), w_j) \quad (4)$$

The gradient vectors in (3) and (4) are computed as follows: for each $n \in \{1,2,\ldots,R_i\}$ and $d \in \{1,2,\ldots,D\}$,

$$\frac{\partial \varphi(\mathbf{x}(t), \omega_i)}{\partial \mu_{n(d)}^i} =$$

$$2P_n^i \left[\sum_{m=1}^{R_i} P_m^i \mathcal{N}_{m,n,d}^i (\mu_{n(d)}^i - \mu_{m(d)}^i) \right.$$

$$\left. - \sum_{m=1}^{R_t} P_m^t \mathcal{N}_{m,n,d}^t (\mu_{m(d)}^t - \mu_{n(d)}^i) \right],$$

and

$$\frac{\partial \varphi(\mathbf{x}(t), \omega_i)}{\partial (\sigma_{n(d)}^i)^2} =$$

$$P_n^i \left[\sum_{m=1}^{R_i} P_m^i \mathcal{N}_{m,n,d}^i \left(\frac{(\mu_{m(d)}^i - \mu_{n(d)}^i)^2}{(\sigma_{m(d)}^i)^2 + (\sigma_{n(d)}^i)^2} - 1 \right) \right.$$

$$\left. - \sum_{m=1}^{R_t} P_m^t \mathcal{N}_{m,n,d}^t \left(\frac{(\mu_{m(d)}^t - \mu_{n(d)}^i)^2}{(\sigma_{m(d)}^t)^2 + (\sigma_{n(d)}^i)^2} - 1 \right) \right],$$

where $\mathcal{N}_{m,n,d}^x = \dfrac{\mathcal{G}(\theta_m^x, \theta_n^i)}{(\sigma_{m(d)}^x)^2 + (\sigma_{n(d)}^i)^2}$.

The next section presents the proposed multi-modal image query and retrieval approach integrating paradigm of text-based and *visual keyword*-based query and retrieval.

3. Multi-modal Image Query and Retrieval

The objective of the proposed multi-modal image query and retrieval (MIQR) method was to make the query result acceptable by seamlessly integrating paradigm of text-based and *visual word*-based query and retrieval. One noticeable peculiarity of the proposed method is to allow a user to query images without consciously segmenting images into objects. More specifically, the proposed method let the user just simply select the reference pixels. Then, the color and texture characteristics of the areas around the reference pixels are collected as partial features of the query objects. All the selected query objects and features can be combined with one another using Boolean operations, such as union, intersection, and exclusion. In other words, image objects are considered as keywords in document search.

A flowchart of the proposed MIQR approach is shown in Figure 4.

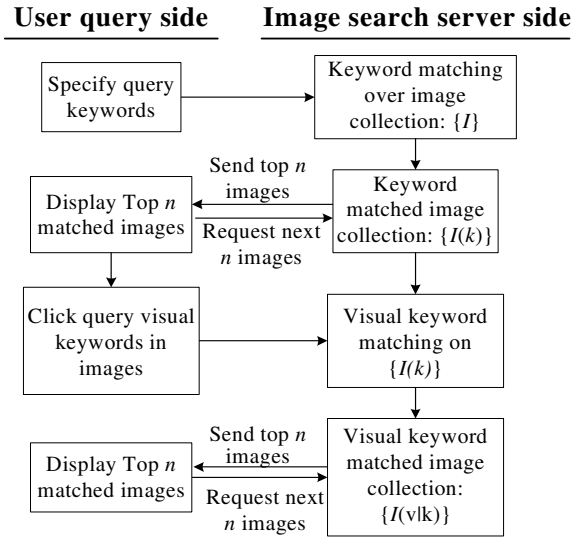
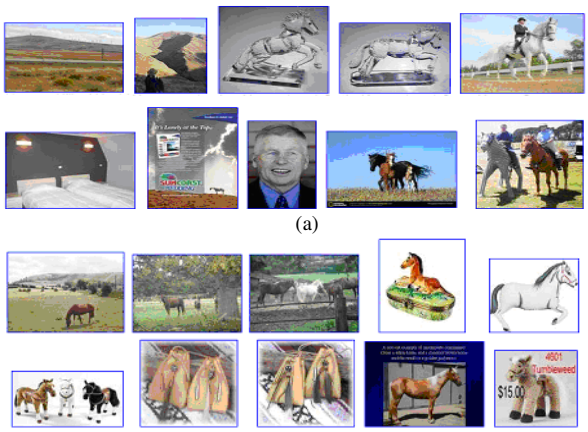


Fig. 4 The flowchart of image query and retrieval procedures for the proposed **MIQR** approach.

First, a user may specify one or several keywords to query and retrieve a set of relevant images from image collections. Figures 5(a) and 5(b) illustrate the retrieval results of the queries on (1) *horse* and (2) *white horse and brown horse*, respectively. As it turned out, neither of these results is satisfactory since the text-based retrieval returns the images containing keywords in their textual contexts instead of keyword-associated information in their visual contents. As we can see, the keyword *horse* is too broad

for the given query topic; likewise, the retrieved images with the keywords *white horse and brown horse* contain white and brown horses, but fewest of them show horses as per user's desired color and/or texture.



(a)

(b)

Fig. 5 Results of text retrieval using keywords (a) *horse* and (b) *white horse and brown horse*.

Therefore, the user may then browse the retrieved images to find horses with desired color and/or texture as query *visual keywords* for secondary search and retrieval. For example, as shown in Figure 6, a user who wants to retrieve pictures presenting *white horse and brown horse* may enter search queries by clicking on a white horse and a brown horse of one or several given sample images. Finally, the retrieval results of the combination of the previous two methods appear in Figure 7. Compared with Figure 5(b), Figure 7 shows a more relevant set of results by prioritizing keywords retrieval matches that are also visually consistent with color and texture features.

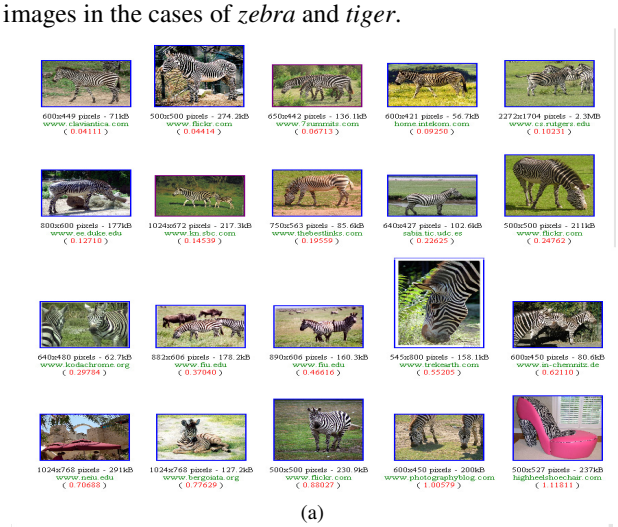
(a) (b)

Fig. 6 Query *visual keywords*: (a) a *brown horse*, and (b) a *white horse*

Fig. 7 Results of visual keywords based retrieval using white and brown colors and textures as the query visual keywords.

4. Experiments

A prototype was constructed to implement the proposed MIQR approach. The experiments were performed on two image sets: (1) a set of 5000 images collected from 10 categories of Corel photo gallery, and (2) a set of 20,000 images randomly gathered from WWW. Thereafter, these two image sets are called Corel5k and WWW20k, respectively. The 10 categories in Corel5k are *butterfly*, *bus*, *elephant*, *flower*, *building*, *dinosaur*, *mountain*, *Africa*, *beach*, and *food*. Each image in WWW20k was initial labeled by the keywords extracted from the named entities in the surrounding text of the image according to the method proposed by Srihari et al. [3] The retrieval examples are illustrated in Figure 8(a) and Figure 8(b). The images at the top-left corner of each figure are the query images, and the rest images are the top 20 query results. As we can see, most of query results match with the query images in the cases of *zebra* and *tiger*.



(a)

(b)

Fig. 8 The retrieval results of the proposed MIQR system by using (a) a *zebra* and (b) *tiger* as the query *visual keywords*.

Two types of experiments were conducted to evaluate the performance of MIQR approach. The first type of experiment intended to show the retrieval performance of MIQR approach on different categories of images. The performance was evaluated according to the averaged image retrieval accuracy versus a sequence of queries, which were applied in the following order: keyword, color, and texture of *visual keywords*. The accuracy [27] is defined as the ratio of relevant images in the top 20 retrieved images. The average accuracy is simply the average of the accuracies measured for the 1600 randomly selected test queries.

Experiment I-1- *queries on a particular object or category*: The query and retrieval experiments were conducted on *butterfly*, *bus*, *elephant*, *flower*, *building*, and *dinosaur* images from Corel5k and WWW20k image collections. The retrieval performance is given in Table 1.

Table 1: Average retrieval rates of Experiment I-1 for searching on Corel5k and WWW20k (values for WWW20k is printed in bold face).

Image Category	Keywords	Color	Texture
butterfly	63/**49**%	70/**56**%	80/**65**%
Bus	35/**23**%	39/**26**%	44/**31**%
elephant	37/**21**%	38/**25**%	41/**30**%
flower	39/**28**%	45/**34**%	55/**43**%
building	28/**21**%	35/**27**%	46/**33**%
dinosaur	86/**37**%	90/**42**%	95/**47**%

Experiment I-2- *query categories without a dominant object or common visual features*: This type of query and retrieval experiments was conducted on *mountain*, *Africa*, *beach*, and *food*, from Corel5k and WWW20k image collections. The retrieval performance is provided in Table 2.

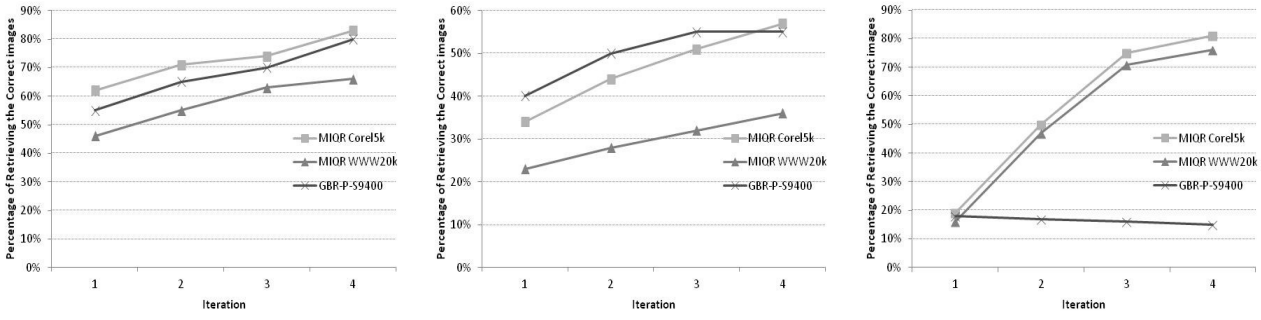
Table 2: Average retrieval rates of Experiment I-2 for searching on Corel5k and WWW20k (values for WWW20k is printed in bold face).

Image Category	Keywords	Color	Texture
mountains	34/**23**%	44/**28**%	54/**34**%
Africa	39/**20**%	45/**29**%	54/**42**%
beach	21/**19**%	25/**22**%	30/**26**%
food	26/**13**%	35/**19**%	47/**27**%

The results on **Experiment I** show that the proposed MIQR approach made the query result acceptable by integrating paradigm of text-based and *visual keyword*-based query and retrieval.

The second type of experiments we carried out was aimed to compare the performance of the MIQR approach and the GBR-P-S approach proposed by Hsu and Li [28]. To evaluate the performance of MIQR approach, we use the average accuracy versus query iterations. The iterations referred to the query sequences as in **Experiment I** for MIQR approach.

Experiment II-1- *query on a particular object or category*: We conducted this experiment by querying and retrieving six categories of images, *butterfly*, *bus*, *elephant*, *flower*, *building*, and *dinosaur* from Corel5k and WWW20k image collections. The retrieval performance of query on butterfly is shown in Figure 9(a). For the query on butterfly, the MIQR achieves 85% and 77% of accuracies over the Corel5k and WWW20k image collection, respectively. By repeating four cycles of user relevant feedback, Hsu [28] achieves an 80% of accuracy on a similar type of query on the butterfly category, with over 9400 images selected from a Corel photo gallery. A more interesting comparison would be made if the MIQR retrieval with comparable image collection were available.



(a) (b) (c)

Fig. 9 Performance comparison of image query and retrieval of (a) butterfly, (b) mountain, and (c) kitchen relevant images among MIQR Corel5k, MIQR WWW20k and GBR-P-S9400. The accuracy is measured as the ratio of relevant images in top 20 retrieved images.

Experiment II-2- *query on background scene*: It is very important for an image query system to have capability to retrieve relevant images according to the query on background scenes. Thus, we conducted experiments to retrieve four different types of images, such as mountain, Africa, beach, and food from both Corel5k and WWW20k image collections. The experimental results of query on mountain are shown in Figure 9(b). Similar to Experiment II-1, MIQR approach is slightly superior to the GBR-P-S approach on the Corel image collection, and is inferior to the GBR-P-S approach on the WWW image collection. Notice that the retrieval accuracy of these experiments do

not imply that the performance of MIQR approach is superior or inferior to the GBR-P-S approach [28] since the MIQR approach uses regions or object-based visual features directly from users' query targets, which usually reflect user's desires.

Experiment II-3- *query on combined objects or regions*: This experiment shows query on a certain category, which has neither dominant regions or objects, nor common visual features. For instance, kitchen can be a representative example in this category. We conducted query and retrieval for kitchen relevant images over both Corel5k and WWW20k image collections. Although a kitchen image has neither dominant regions nor common visual features, a kitchen in the real world usually contains several different kitchen utensils and furniture, such as a microwave oven, a refrigerator, a dishwasher, kitchen chairs, etc. Thus, the MIQR approach allows a user to select a few kitchen objects or regions as user's queries for retrieving images associated with kitchen category. The GBR-P-S approach achieves 18% of accuracy on the kitchen category retrieval over their Corel image collection. As shown in Figures 9(c), MIQR approach achieves 75% and 70% of accuracies from querying and retrieving the Corel5k and WWW20k image collections, respectively.

The results on the **Experiment II** show that the proposed MIQR approach achieved comparable accuracy to the current leading approaches [28]. To summarize the results of the experiments, it appears that (1) the MIQR approach can retrieve relevant images closely associated with users' query regions or objects, (2) the MIQR approach has the capability of searching and retrieval relevant images from a large collection such as the WWW20k, and (3) the MIQR approach allows users to combine several *visual keywords* to organize a category, which has neither dominant objects nor common visual features, such as kitchen, laboratory, etc.

5. Conclusions

In this paper, we introduce a new multi-modal image query and retrieval (MIQR) approach based on {\it visual keywords} and GPDNN. A comprehensive performance evaluation of the proposed MIQR approach is given using a subset of the COREL image gallery and the images gathered from WWW, and a comparison with current leading approaches [28] is made. The experimental results indicate that the proposed MIQR approach made the query result acceptable by integrating paradigm of text-based and *visual keyword*-based query and retrieval, and achieve comparable accuracy to those of Hsu [28]. Therefore, we can conclude that the proposed MIQR approach (1) can retrieve relevant images closely associated with users' query regions or objects, (2) has the capability of searching and retrieval relevant images from a large collection, and (3) allows users to combine several queries to organize a category, which has neither dominant regions or objects, nor common visual features, such as kitchen, laboratory, etc. Among the images collected from WWW, we see quite a few images are illustrated for the representation or summary of their associated videos. Therefore, a user may carefully select search queries to retrieve desired images and associated videos from WWW. Future work will hopefully collect more images from the WWW to let users be able to query and retrieval more desired images and videos.

Acknowledgments

This work was supported by the National Science Council, Taiwan, R.O.C., under Grant No.NSC100-2221-E-241-014.

References

[1] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys, Vol. 40, No. 2, 2008, pp. 5:1-5:60
[2] H. T. Shen, B. C. Ooi, K.-L. Tan, Giving meanings to www images, in: MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia, New York, NY, USA, 2000, pp. 39-47.
[3] R. K. Srihari, Z. Zhang, A. Rao, H. Baird, F. Chen, Intelligent indexing and semantic retrieval of multimodal documents, Information Retrieval 2, 1999, pp. 245-275.
[4] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, 2000, pp. 888-905.
[5] W. Y. Ma, B. S. Manjunath, Edgeflow: a technique for boundary detection and image segmentation, IEEE Transactions on Image Processing, Vol. 9, No. 8, 2000, pp. 1375-1388.
[6] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, Unsupervised multiresolution segmentation for images with low depth of field, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 1, 2001, pp. 85-90.
[7] J. Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: Semantics-sensitive integrated matching for picture LIbraries, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 9, 2001, pp. 947-963.
[8] Y. Chen, J. Z. Wang, A region-based fuzzy feature matching approach to content-based image retrieval, IEEE Transactions on Pattern analysis and machine intelligent, Vol. 24, No. 9 , 2002, pp. 1252-1267.
[9] V. H. Kondekar, V. S. Kolkure, S. N. Kore, Image retrieval techniques based on image features, a state of art approach for CBIR, International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010, pp. 69-76.
[10] X. Wang, Y. J. Yu, H. Y. Yang, An effective image retrieval scheme using color, texture and shape features, Computer Standards & Interfaces, Vol. 33, No. 1, 2011, pp. 59-68.
[11] A. Pentland, R. Picard, S. Sclaroff, Photobook: Content-based manipulation of image databases, in: Storage and

Retrieval for Image, Video Databases II, Vol. 2185, SPIE, San Jose, CA, 1994.
[12] J. R. Smith, S. F. Chang, VisualSEEk: a fully automated content-based image query system, in: Proceedings of the fourth ACM international conference on Multimedia, ACM, New York, NY, USA, 1996, pp. 87-98.
[13] W. Y. Ma, B. S. Manjunath, Netra: A toolbox for navigating large image databases, in: Proc. IEEE Int'l Conf. Image Processing, 1997, pp. 568-571.
[14] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, Blobworld: A system for region-based image indexing and retrieval, in: Third International Conference on Visual Information Systems, 1999, pp. 509-516.
[15] W. Huang, Y. Gao, K. L. Chan, A review of region-based image retrieval, Signal Processing Systems, Vol. 59, No. 2, 2010, pp. 143-161.
[16] J. Vogel, B. Schiele, Semantic modeling of natural scenes for content-based image retrieval, International Journal of Computer Vision, Vol. 72, No. 2, 2007, pp. 133-157.
[17] N. Vasconcelos, From pixels to semantic spaces: Advances in content-based image retrieval, IEEE Computer, Vol. 40, No. 7, 2008, pp. 20-26.
[18] J. Yu, Q. Tian, Semantic subspace projection and its applications in image retrieval, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, No. 4, 2008, pp. 544-548.
[19] T. M. Deserno, S. Antani, L. R. Long, Ontology of gaps in content-based image retrieval, Journal of Digital Imaging, Vol. 22, No. 2, 2009, pp. 202-215.
[20] K. Goh, B. Li, E. Y. Chang, Semantics and feature discovery via confidence-based ensemble, TOMCCAP 1 (2) , 2005, pp. 168-189.
[21] H. Xu, H. Yaguchi, Visual evaluation at scale of threshold to suprathreshold color difference, Color Research and: Application, Vol. 30, No. 3, 2005, pp. 198-208.
[22] J. G. Daugman, Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 36, No. 7 , 1988, pp.1169-1179.
[23] B. S. Manjuath, W. Y. Ma, Texture features for browsing and retrieval of image data, IEEE Transactions on PAMI, Vol. 18, No. 8, 1996, pp. 837-842.
[24] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, 1977, pp. 1-38.
[25] Y. Y. Xu, S. C. Chuang, C. L. Tseng, H. C. Fu, Generalized probabilistic decision-based neural networks for texture classification and retrieval, in: The 2010 International Conference on Modelling, Identification and Control (ICMIC), Okayama, Japan, 2010, pp. 500-504.
[26] H. C. Fu, Y. Y. Xu, Multilinguistic handwritten character recognition by Bayesian decision-based neural networks, IEEE Transactions on Signal Processing, Vol. 46, No. 10, 1998, pp. 2781-2789.
[27] Z. Su, H. Zhang, S. Z. Li, S. Ma, Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning, IEEE Transactions on Image Processing, Vol. 12, No. 8, 2003, pp. 924-937.
[28] C. T. Hsu, C. Y. Li, Relevance feedback using generalized Bayesian framework with region-based optimization learning, IEEE Transactions on Image Processing, Vol. 14, No. 10, 2005, pp. 1617-1631.

Yeong-Yuh Xu received his B.S. degree in electrical engineering in 1995 from National Sun Yat-Sen University, Kaohsiung, Taiwan. He received his M.S. and Ph.D. degree in computer science and information engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2004, respectively. From 2005 to 2009, he served as a Postdoctoral Fellow in the Department of Computer Science and Information Engineering of National Chiao-Tung University, Hsinchu, Taiwan. Currently, he is an assistant professor in the Department of Computer Science and Information Engineering, Hungkuang University, Taichung, Taiwan. Dr. Xu's research interests include pattern recognition, neural networks, and content-based image/video retrieval.

Hsin-Chia Fu received the B.S. degree from National Chiao-Tung University in Electrical and Communication engineering in 1972, and the M.S. and Ph.D. degrees from New Mexico State University, both in Electrical and Computer Engineering in 1975 and 1981, respectively. From 1981 to 1983 he was a Member of the Technical Staff at Bell Laboratories. Since 1983, he has been on the faculty of the Department of Computer science and Information engineering at National Chiao-Tung University, in Taiwan, ROC. From 1987 to 1988, he served as the director of the department of information management at the Research Development and Evaluation Commission, of the Executive Yuan, ROC. From 1988-1989, he was a visiting scholar of Princeton University. From 1989 to 1991, he served as the chairman of the Department of Computer Science and Information Engineering. From September to December of 1994, he was a visiting scientist at Fraunhofer-Institut for Production Systems and Design Technology (IPK), Berlin Germany. His research interests include digital signal/image processing, VLSI array processors, and neural networks. Dr. Fu was the co-recipient of the 1992 and 1993 Long-Term Best Thesis Award with Koun Tem Sun and Cheng Chin Chiang, and the recipient of the 1996 Xerox OA paper Award. He has served as a founding member, Program co-chair (1993) and General co-chair (1995) of International Symposium on Artificial Neural Networks. He is presently serving on Technical Committee on Neural Networks for Signal Processing of the IEEE Signal Processing Society. He has authored more than 100 technical papers, and two textbooks ``PC/XT BIOS Analysis'', and ``Introduction to neural networks'', by Sun-Kung Book Co., and Third Wave Publishing Co., respectively. Dr. Fu is a member of the IEEE Signal Processing and Computer Societies, Phi Tau Phi, and the Eta Kappa Nu Electrical Engineering Honor Society.