Pattern Recognition 43 (2010) 2380–2389

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image retrieval based on multi-texton histogram

Guang-Hai Liu [a,*], Lei Zhang [b], Ying-Kun Hou [d], Zuo-Yong Li [c], Jing-Yu Yang [c]

[a] College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[c] Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China
[d] School of Information Science and Technology, Taishan University, Taian 271021, China

ARTICLE INFO

Article history:
Received 30 March 2009
Received in revised form
16 January 2010
Accepted 11 February 2010

Keywords:
Image retrieval
Texton detection
Multi-texton histogram

ABSTRACT

This paper presents a novel image feature representation method, called multi-texton histogram (MTH), for image retrieval. MTH integrates the advantages of co-occurrence matrix and histogram by representing the attribute of co-occurrence matrix using histogram. It can be considered as a generalized visual attribute descriptor but without any image segmentation or model training. The proposed MTH method is based on Julesz's textons theory, and it works directly on natural images as a shape descriptor. Meanwhile, it can be used as a color texture descriptor and leads to good performance. The proposed MTH method is extensively tested on the Corel dataset with 15 000 natural images. The results demonstrate that it is much more efficient than representative image feature descriptors, such as the edge orientation auto-correlogram and the texton co-occurrence matrix. It has good discrimination power of color, texture and shape features.

1. Introduction

Image retrieval is an important topic in the field of pattern recognition and artificial intelligence. Generally speaking, there are three categories of image retrieval methods: text-based, content-based and semantic-based. The text-based approach can be traced back to 1970s [4]. Since the images need to be manually annotated by text descriptors, it requires much human labour for annotation, and the annotation accuracy is subject to human perception. In early 1990s, researchers had built many content-based image retrieval systems, such as QIBC, MARS, Virage, Photobook, FIDS, Web Seek, Netra, Cortina [5], VisualSEEK [6] and SIMPLIcity [7]. Various low-level visual features can be extracted from the images and stored as image indexes. The query is an image example that is indexed by its features, and the retrieved images are ranked with respect to their similarity to the query image. Since the indexes are directly derived from the image content, it requires no semantic labeling [8]. Considering that humans tend to use high-level features to interpret images and measure their similarity and image low-level features (e.g. color, texture, shape) often fail to describe the high level semantic concepts, researchers have proposed some methods for image retrieval by using machine learning techniques such as SVM [9–13].

Some statistical models have been proposed to exploit the similarities between image regions or patches, which are represented in a uniform vector, such as the Visual Token Catalog [14] and the Visual Language Modeling [15]. They map the blob to visual words and apply language model to visual words. A visual token catalog is generated to exploit the content similarities between regions in [14], while the Visual Language Modeling in [15] is based on the assumption that there are implicit visual grammars in a meaningful image. Those methods need accurate image segmentation, which is however still an open problem. Limited by the current advances of artificial intelligence and cognitive science, semantic-based image retrieval still has a long way to go for real applications. Comparatively, content-based image retrieval (CBIR) is still attracting much attention by researchers. In general, the research of CBIR techniques mainly focuses on two aspects: part-based object retrieval [16–19] and low-level visual feature-based image retrieval [20–23].

In this paper, we focus on edge-based image representation for image retrieval. In [24,25], Jain et al. introduced the edge direction histogram (EDH) for trademark images retrieval. This method is invariant to image translation, rotation and scaling because it uses the edges only but ignores correlation between neighboring edges. EDH only suits for flat-images of trademarks. Gevers et al. [23,35] proposed a new method for image indexing and retrieval by combining color and shape invariant features. This method is robust to partial occlusion, object clutter and change in viewpoint. The MPEG-7 edge histogram descriptor (EHD) can capture the spatial distribution of edges, and it is an efficient texture descriptor for images with heavy textural presence. It can also

* Corresponding author. Tel./fax: +86 25 84315510.
E-mail addresses: liuguanghai009@163.com (G.-H. Liu), cslzhang@comp.polyu.edu.hk (L. Zhang), yangjy@mail.njust.edu.cn (J.-Y. Yang).

0031-3203/$ - see front matter Crown Copyright © 2010 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.012

work as a shape descriptor as long as the edge field contains the true object boundaries [26]. In [27], Mahmoudi et al. proposed the edge orientation autocorrelogram (EOAC) for shape-based image indexing and retrieval. It can be used for edge-based image indexing and retrieval without segmentation. The EOAC is invariant to translation, scaling, color, illumination, and small viewpoint variations, but it is not appropriate for texture-based images retrieval. Lowe [28] proposed a very effective algorithm, called scale-invariant feature transform (SIFT), in computer vision to detect and describe local features in images. It has been widely used in object recognition, robotic mapping and navigation, image stitching, 3D modeling, gesture recognition, video tracking, etc. Banerjee et al. [31] proposed to use edge-based features for CBIR. The algorithm is computationally attractive as it computes different features with limited number of selected pixels. The texton co-occurrence matrices (TCM) proposed in [20] can describe the spatial correlation of textons for image retrieval. It has the discrimination power of color, texture and shape features. Kiranyaz et al. [21] proposed a generic shape and texture descriptor over multi-scale edge field for image retrieval, which is the so called 2-D walking ant histogram (2D-WAH). As a shape descriptor, it deals directly with natural images without any segmentation or object extraction preprocessing stage. When tuned as a texture descriptor, it can achieve good retrieval accuracy especially for directional textures. Luo et al. [38] developed a robust algorithm called color edge co-occurrence histogram (CECH), which is based on a particular type of spatial-color joint histogram. This algorithm employs perceptual color naming to handle color variation, and pre-screening to limit the search scope (i.e. size and location) of the object.

Natural scenes are usually rich in both color and texture, and a wide range of natural images can be considered as a mosaic of regions with different colors and textures. The human visual system exhibits a remarkable ability to detect subtle differences in textures that are generated from an aggregate of fundamental micro-structures or elements [1,2]. Color and texture have close relationship via fundamental micro-structures in natural images and they are considered as the atoms for pre-attentive human visual perception. The term "texton" is conceptually proposed by Julesz [1]. It is a very useful concept in texture analysis and has been utilized to develop efficient models in the context of texture recognition or object recognition [33,34]. However, few works were proposed to apply texton models to image retrieval. How to obtain texton features, and how to map the low-level texture features to textons need to be further studied. To this end, in this paper we propose a new descriptor for image retrieval. It can represent the spatial correlation of color and texture orientation without image segmentation and learning processes.

This paper presents a new feature extractor and descriptor, namely multi-texton histogram (MTH), for image retrieval. MTH can be viewed as an improved version of TCM. It is specially designed for natural image analysis and can achieve higher retrieval precision than that of EOAC [27] and TCM [20]. It integrates the advantages of co-occurrence matrix and histogram by representing the attribute of co-occurrence matrix using histogram, and can represent the spatial correlation of color and texture orientation.

The rest of this paper is organized as follows. In Section 2, the TCM is introduced. The MTH is presented in Section 3. In Section 4, performance comparison among EOAC, TCM and MTH is taken on two Corel datasets. Section 5 concludes the paper.

2. The texton co-occurrence matrix (TCM)

Before describing in detail the proposed MTH, let us briefly review the TCM [20] method for image retrieval. TCM can represent the spatial correlation of textons, and it can discriminate color, texture and shape features simultaneously. Let r, g and b be unit vectors along the R, G and B axes in RGB color space, we define the following vectors for a full color image $f(x,y)$ [3,32]:

$$u = \frac{\partial R}{\partial x} r + \frac{\partial G}{\partial x} g + \frac{\partial B}{\partial x} b \tag{1}$$

$$v = \frac{\partial R}{\partial y} r + \frac{\partial G}{\partial y} g + \frac{\partial B}{\partial y} b \tag{2}$$

g_{xx}, g_{yy} and g_{xy} are defined as the dot products of these vectors:

$$g_{xx} = u^T u = |\partial R/\partial x|^2 + |\partial G/\partial x|^2 + |\partial B/\partial x|^2 \tag{3}$$

$$g_{yy} = v^T v = |\partial R/\partial y|^2 + |\partial G/\partial y|^2 + |\partial B/\partial y|^2 \tag{4}$$

$$g_{xy} = u^T v = \frac{\partial R}{\partial x}\frac{\partial R}{\partial y} + \frac{\partial G}{\partial x}\frac{\partial G}{\partial y} + \frac{\partial B}{\partial x}\frac{\partial B}{\partial y} \tag{5}$$

Let $v(x,y)$ be an arbitrary vector in RGB color space. Using the above notations, it can be seen that the direction of maximum rate of the change of $v(x,y)$ is [3,32]

$$\theta(x,y) = \frac{1}{2}\tan^{-1}\left[\frac{2g_{xy}}{(g_{xx}-g_{yy})}\right] \tag{6}$$

The value of the rate of change at (x, y) in the direction of $\theta(x, y)$ is given by

$$G(x,y) = \left\{\tfrac{1}{2}[(g_{xx}+g_{yy})+(g_{xx}-g_{yy})\cos 2\theta + 2g_{xy}\sin 2\theta]\right\}^{1/2} \tag{7}$$

Denote by Max(G) and Min(G) the maximum and minimum values of G along some direction by Eq. (7). The original color image is quantized into 256 colors in RGB color space, denoted by $C(x, y)$. Five special types of texton templates are used to detect the textons, which are shown in Fig. 1. The flow chart of texton detection is illustrated in Fig. 2. In an image, we move the 2×2 grid from left-to-right and top-to-bottom throughout the image to detect textons with one pixel as the step-length. If the pixel values that fall in the texton template are the same, those pixels will form a texton, and their values are kept as the original values. Otherwise they will be set to zero. Each texton template can lead to a texton image (an example of texton detection result is shown in Fig. 2(a)), and the five texton templates will lead to five texton images. We combine them into a final texton image, as shown in Fig. 2(b).

For the texton images detected with Max(G), Min(G) and $C(x, y)$, we use co-occurrence matrix to extract their features. Denote the values of a texton image as $f(P)=w$, $w \in \{0, 1, \ldots, 255\}$. The pixel position is $P=(x, y)$. Let $P_1=(x_1, y_1)$, $P_2=(x_2, y_2)$, $f(P_1)=w$ and $f(P_2)=\hat{w}$. If the probability of two values w and $\hat{w}$ co-occur with two pixel positions related by D, we define the cell entry $(w, \hat{w})$ of co-occurrence matrix $C_{D,\theta}$ as follows:

$$C_{D,\theta} = 1 - Pr\{f(P_1) = w \wedge f(P_2) = \hat{w} || P_1 - P_2 | = D\} \tag{8}$$

The TCM utilizes energy, contrast, entropy and homogeneity to describe image features. For an image, a 12-dimensional vector will be obtained as the final feature for retrieval.

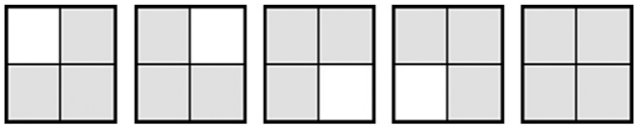

Fig. 1. Five special texton types used in TCM.

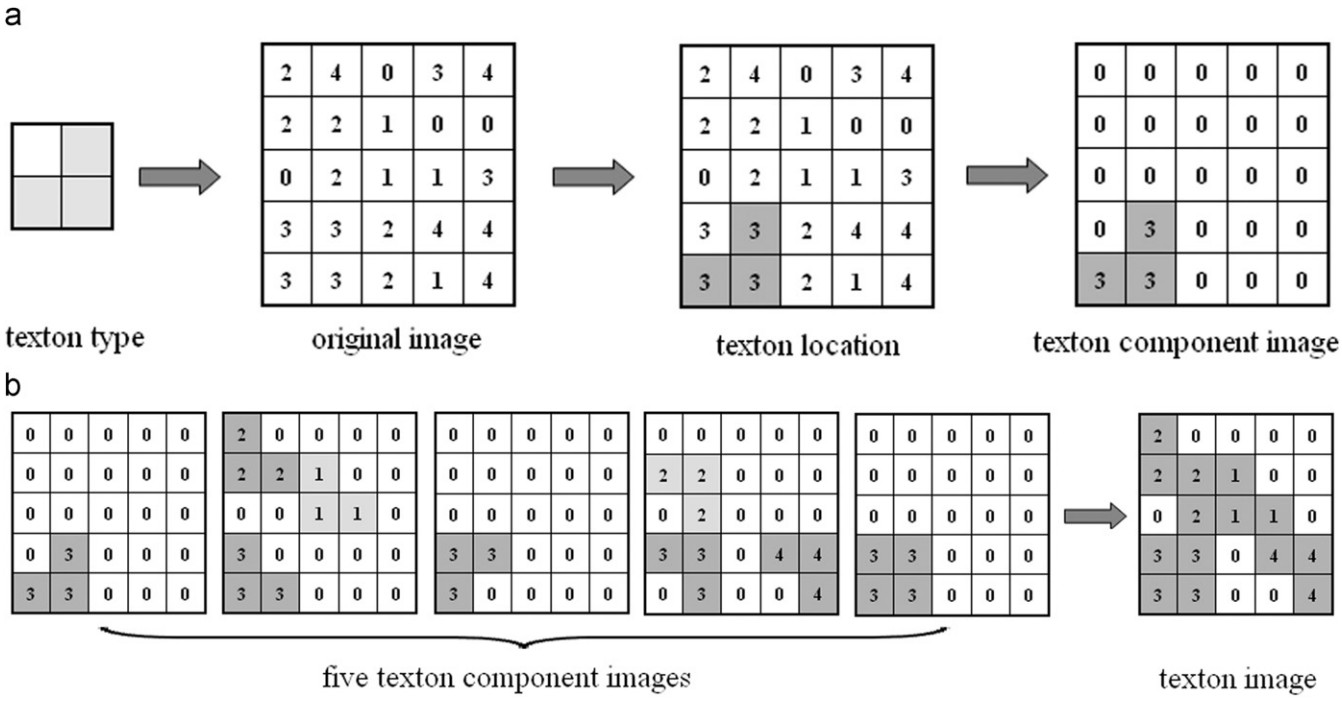

Fig. 2. The flow chart of texton detection in TCM: (a) an example of texton detection; (b) the five detected texton images and the final texton image.

3. The multi-texton histogram (MTH)

The study of pre-attentive (also called effortless) texture discrimination can serve as a model system, with which the roles of local texture detection and global (statistical) computation in visual perception can be distinguished [1]. This can be easily explained by the local orientation differences between the elements that constitute the two texture images. It is possible to describe the differences between the texture images globally. If the first-order statistics of two texture images are identical, the second-order statistics may also differ greatly [1]. The first and second-order statistics have their own advantages in texture discrimination, so in this paper we propose to combine the first-order statistics and second-order statistics into an entity for texton analysis. We call the proposed technique multi-texton histogram (MTH), and use it for image retrieval.

Based on the texton theory [1,2], texture can be decomposed into elementary units, the texton classes of colors, elongated blobs of specific widths, orientation and aspect ratios, and the terminators of these elongated blobs. In the proposed MTH-based image retrieval scheme, texture orientation needs to be detected for texton analysis. In the following sub-section, we propose a computationally efficient method for texture orientation detection.

3.1. Texture orientation detection

Texture orientation analysis plays an important role in computer vision and pattern recognition. For instance, orientation is used in pre-attentive vision to characterize textons [1–3]. Orientation of texture images has a strong influence on human's perception of a texture image. Texture orientation can also be used to estimate the shape of textured images. The orientation map in an image represents the object boundaries and texture structures, and it provides most of the semantic information in the image. In this paper, we propose a computationally efficient algorithm for texture orientation detection.

By applying some gradient operator, such as the Sobel operator, to a gray level image along horizontal and vertical directions, we can have two gradient images, denoted by g_x and g_y. A gradient map $g(x,y)$ can be obtained, with the gradient magnitude and orientation defined as $|g(x,y)|=\sqrt{g_x^2+g_y^2}$ and $\theta(x,y)=\arctan(g_y/g_x)$.

As for full color images, there are red, green and blue channels. If we convert the full color image into a gray image, and then detect the gradient magnitude and orientation from the gray image, much chromatic information will lose. In order to detect the edges caused by chromatic changes, we propose the following method.

In the Cartesian space, let $a=(x_1,y_1,z_1)$ and $b=(x_2,y_2,z_2)$. Their dot product is defined as

$$a\cdot b=x_1x_2+y_1y_2+z_1z_2 \tag{9}$$

so that

$$\cos(\widehat{a,b})=\frac{a.b}{|a||b|}=\frac{x_1x_2+y_1y_2+z_1z_2}{\sqrt{x_1^2+y_1^2+z_1^2}\cdot\sqrt{x_2^2+y_2^2+z_2^2}} \tag{10}$$

We apply the Sobel operator to each of the red, green and blue channels of a color image $f(x,y)$. The reason that we use the Sobel operator is that it is less sensitive to noise than other gradient operators or edge detectors while being very efficient [3]. The gradients along x and y directions can then be denoted by two vectors $a(R_x,G_x,B_x)$ and $b(R_y,G_y,B_y)$, where R_x denotes the gradient in R channel along horizontal direction, and so on. Their norm and dot product can be defined as

$$|a|=\sqrt{(R_x)^2+(G_x)^2+(B_x)^2} \tag{11}$$

$$|b|=\sqrt{(R_y)^2+(G_y)^2+(B_y)^2} \tag{12}$$

$$a\cdot b=R_x\cdot R_y+G_x\cdot G_y+B_x\cdot B_y \tag{13}$$

The angle between a and b is then

$$\cos(\widehat{a,b})=\frac{a\cdot b}{|a|\cdot|b|} \tag{14}$$

$$\theta = \arccos[\cos(\widehat{a,b})] = \arccos\left[\frac{a \cdot b}{|a| \cdot |b|}\right] \quad (15)$$

After the texture orientation θ of each pixel is computed, we quantize it uniformly into 18 orientations with 10° as the step-length.

3.2. Color quantization in RGB color space

It is well known that color provides powerful information for image retrieval or object recognition, even in the total absence of shape information. HSV color space could mimic human color perception well, and thus many researchers use it for color quantization. In terms of digital processing, however, RGB color space is most commonly used in practice and it is straightforward. In order to extract color information and simplify manipulation, in this work the RGB color space is used and it quantized into 64 colors. In Section 4.4, the experiments demonstrated that the RGB color space is well suitable for our framework. Given a color image with size $N \times N$, we uniformly quantize the R, G, and B channels into 4 bins so that 64 colors are obtained. Denote by $C(x, y)$ the quantized image, where $x, y = [0, 1, \ldots, N-1]$. Then each value of $C(x, y)$ is a 6-bits binary code, ranging from 0 to 63.

3.3. Texton detection

The concept of "texton" was proposed in [1] more than 20 years ago, and it is a very useful tool in texture analysis. In general, textons are defined as a set of blobs or emergent patterns sharing a common property all over the image; however, defining textons remains a challenge. In [2], Julesz presented a more complete version of texton theory, with emphasis on critical distances (Δ) between texture elements on which the computation of texton gradients depends. Textures are formed only if the adjacent elements lie within the Δ-neighborhood. However, this Δ-neighborhood depends on element size. If the texture elements are greatly expanded in one orientation, pre-attentive discrimination is somewhat reduced. If the elongated elements are not jittered in orientation, this increases the texton-gradients at the texture boundaries. Thus, with a small element size, such as 2×2, texture discrimination can be increased because the texton gradients exist only at texture boundaries [2]. In view of this and for the convenience of expression, the 2×2 block is used in this paper for textons detection.

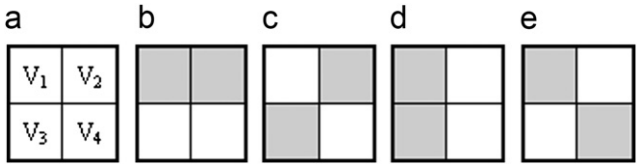

Fig. 3. Four texton types defined in MTH: (a) 2×2 grid; (b) T_1; (c) T_2; (d) T_3 and (e) T_4.

The texton templates defined in MTH are different from those in TCM (refer to Fig. 1). In this paper, four special texton types are defined on a 2×2 grid, as shown in Fig. 3. Denote the four pixels as V_1, V_2, V_3 and V_4. If the two pixels highlighted in gray color have the same value, the grid will form a texton. Those 4 texton types are denoted as T_1, T_2, T_3 and T_4, respectively.

The working mechanism of texton detection is illustrated in Fig. 4. In the color index image $C(x, y)$, we move the 2×2 block from left-to-right and top-to-bottom throughout the image to detect textons with 2 pixels as the step-length. If a texton is detected, the original pixel values in the 2×2 grids are kept unchanged. Otherwise it will have zero value. Finally, we will obtain a texton image, denoted by $T(x, y)$.

The four texton types used in MTH contain richer information than those in TCM because the co-occurring probability of two same-valued pixels is bigger than that of three or four same-valued pixels in a 2×2 grid. As for the texton detection procedure, MTH is also faster than TCM. In the texton detection of TCM, the 2×2 grid moves throughout the image with one pixel as the step-length, and the detected textons in a neighborhood may overlap. The final texton image needs to be fused by the overlapped components of textons, and this will increase the computational complexity. Therefore, in this paper the step-length is set to two pixels to reduce the computational cost.

3.4. Features representation

In [16], the angle and radius are quantized by using the log polar quantization scheme as in [29,30]. The angle is quantized into 12 bins and the radius is quantized into 5 bins. The log-polar quantization has a good performance in image retrieval. It can well express the local information, but the feature dimension is big and the feature matrix is sparse. The TCM scheme utilizes energy, contrast, entropy and homogeneity to describe image features [20]. However, these metrics cannot fully represent the discrimination power of color, texture and shape features. There is still much room to improve TCM, and the proposed method in this section is such an improved version of TCM.

The co-occurrence matrix characterizes the relationship between the values of neighboring pixels, while the histogram-based techniques have high indexing performance and are simple to compute. If we use the co-occurrence matrix to represent image features directly, the dimension will be high and the performance can be decreased. If we use histogram only to represent image features, the spatial information will be lost. In

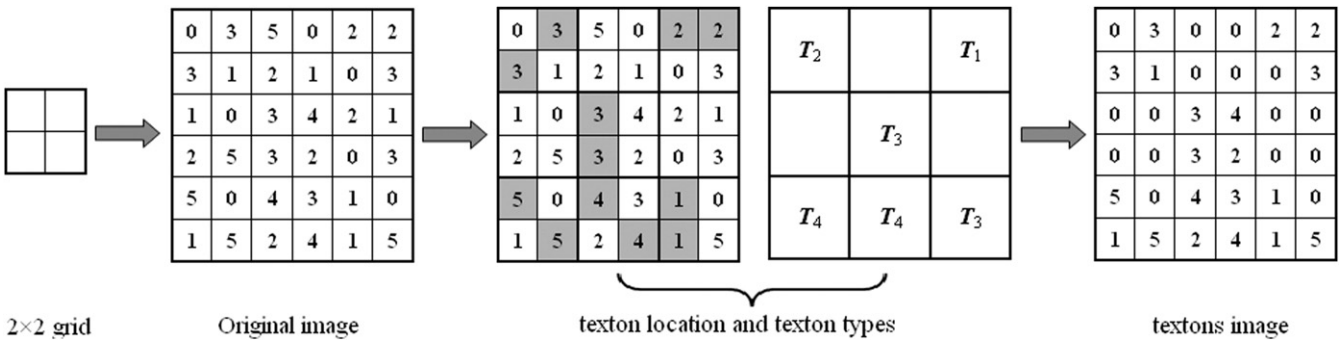

Fig. 4. Illustration of the texton detection process.

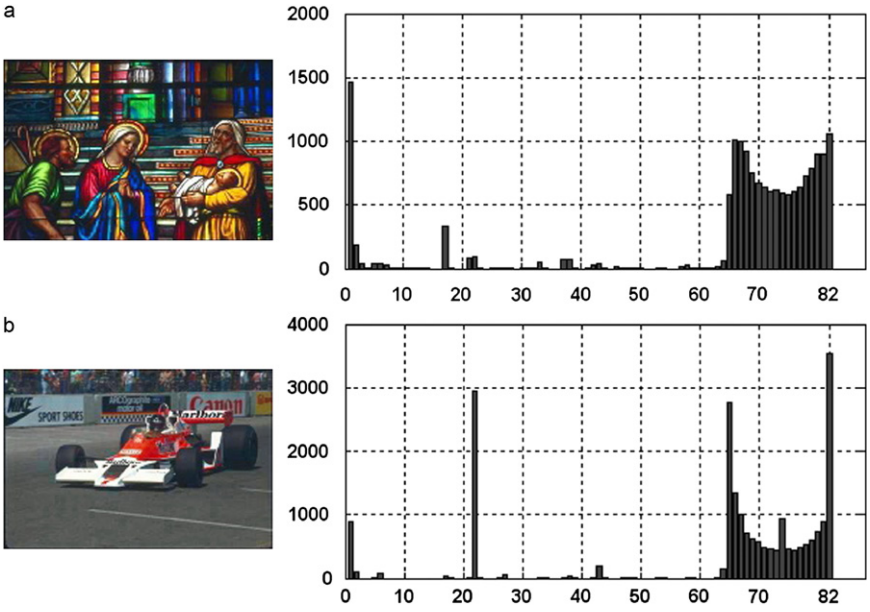

Fig. 5. Two examples of MTH: (a) stained glass; (b) racing car.

order to combine the advantages of co-occurrence matrix and histogram, in this paper we propose the MTH descriptor.

The values of a texton image T are denoted as $w \in \{0, 1, \ldots, W-1\}$. Denote by $P_1=(x_1, y_1)$ and $P_2=(x_2, y_2)$ two neighboring pixels, and their values are $T(P_1)=w_1$ and $T(P_2)=w_2$. In the texture orientation image $\theta(x,y)$, the angles at P_1 and P_2 are denoted by $\theta(P_1)=v_1$ and $\theta(P_2)=v_2$. In texton image T, two different texture orientations may have the same color, while in texture orientation image $\theta(x,y)$, two different colors may have the same texture orientation. Denote by N the co-occurring number of two values v_1 and v_2, and by $\overline{N}$ the co-occurring number of two values w_1 and w_2. With two neighboring pixels whose distance is D, we define the MTH as follows:

$$H(T(P_1)) = \begin{cases} N\{\theta(P_1)=v_1 \wedge \theta(P_2)=v_2 || P_1-P_2 |=D\} \\ \text{where} \quad \theta(P_1)=\theta(P_2)=v_1=v_2 \end{cases} \tag{16}$$

$$H(\theta(P_1)) = \begin{cases} \overline{N}\{T(P_1)=w_1 \wedge T(P_2)=w_2 || P_1-P_2 |=D\} \\ \text{where} \quad T(P_1)=T(P_2)=w_1=w_2 \end{cases} \tag{17}$$

The proposed algorithm analyzes the spatial correlation between neighboring color and edge orientation based on four special texton types, and then forms the textons co-occurrence matrix and describes the attribute of texton co-occurrence matrix using histogram. This is why we call it multi-texton histogram (MTH). Fig. 5 shows two examples of the proposed MTH.

$H(T(P_1))$ can represent the spatial correlation between neighboring texture orientation by using color information, leading to a 64 dimensional vector. $H(\theta(P_1))$ can represent the spatial correlation between neighboring colors by using the texture orientation information, leading to a 18 dimensional vector. Thus in total MTH uses a 64+18=82 dimensional vector as the final image features in image retrieval.

4. Experimental results

In this section, we demonstrate the performance of our method using two Corel datasets. The methods used in comparison are the edge orientation autocorrelogram (EOAC) [27] and TCM [20]. Both EOAC and the proposed MTH are based on edge features without image segmentation, and TCM is the origin of our method. In the experiments, we selected randomly 50 images from every category as query image. The performance is evaluated by the average results of each query respectively. The source code of the proposed MTH algorithm can be downloaded at http://www.comp.polyu.edu.hk/~cslzhang/code/MTH_C_Code.txt.

4.1. Datasets

There are so far no standard test datasets and performance evaluation models for CBIR systems [4]. Although many image datasets, such as Coil-100 dataset, ETH-80 dataset and VisTex texture dataset, are available, they are mainly used for image classification or object recognition. There are essential differences between image retrieval and image classification. Image classification has the training dataset and aims at identifying the class of the query image; however, in image retrieval there is no training set, and the purpose is to search for similar images to the given one. The image datasets used for image classification are not well fitted for image retrieval, and the image representations used in image classification are often not well fitted for image retrieval, either. The Corel image dataset is the most commonly used dataset to test image retrieval performance and the Brodatz texture dataset [36] and the OUTex texture dataset [37] are also widely used. Images collected from internet serve as another data source especially for systems targeting at Web image retrieval [37].

The Corel image database contains a large amount of images of various contents ranging from animals and outdoor sports to natural scenarios. Two Corel datasets are used in our image retrieval systems. The first one is the Corel 5000 dataset, which contains 50 categories. There are 5000 images from diverse contents such as fireworks, bark, microscopic, tile, food texture, tree, wave, pills and stained glass. Every category contains 100 images of size 192×128 in JPEG format. The second dataset is Corel 10 000 dataset. It contains 100 categories. There are 10 000 images from diverse contents such as sunset, beach, flower,

building, car, horses, mountains, fish, food, door, etc. Every category contains 100 images of size 192×128 in JPEG format. The Corel 10 000 dataset contains all categories of Corel 5000 dataset.

4.2. Distance metric

For each template image in the dataset, an M-dimensional feature vector $T=[T_1, T_2 \ldots T_M]$ will be extracted and stored in the database. Let $Q=[Q_1, Q_2 \ldots Q_M]$ be the feature vector of a query image, the distance metric between them is simply calculated as

$$D(T,Q)=\sum_{i=1}^{M}\frac{|T_i-Q_i|}{1+T_i+Q_i} \tag{18}$$

The above formula is as simple to calculate as the L_1 distance, which needs no square or square root operations. It can save much computational cost and is very suitable for large scale image datasets. Actually, it can be considered as a weight L_1 distance with the $1/(1+T_i+Q_i)$ being the weight. For the proposed MTH, $M=82$ for color images. The class label of the template image which yields the smallest distance will be assigned to the query image.

4.3. Performance measure

In order to evaluate the effectiveness of our method, the *Precision* and *Recall* curves are adopted, which are the most common measurements used for evaluating image retrieval performance. *Precision* and *Recall* are defined as follows:

$$P(N)=I_N/N \tag{19}$$

$$R(N)=I_N/M \tag{20}$$

where I_N is the number of images retrieved in the top N positions that are similar to the query image, M is the total number of images in the database similar to the query, and N is the total number of images retrieved. In our image retrieval system, $N=12$ and $M=100$.

4.4. Retrieval performance

In the experiments, different quantization levels of texture orientation and color are used to test the performance of the proposed MTH in RGB color space. The HSV color space is also used for comparison. Denote by $bin(H)$, $bin(S)$ and $bin(V)$ the number of bins for H, S and V components. Similar to [26,39], in this paper we let $bin(H)\geq 8$, $bin(S)\geq 3$ and $bin(V)\geq 3$ for HSV color space quantization in the image retrieval experiments, and hence the number of total bins is at least 72 and it is gradually increased to 128 bins. Tables 1 and 2 provide the average retrieval precision and recall of MTH in both RGB and HSV color spaces. We can see that under the same or similar retrieval precision, the performance of MTH in RGB color space is better than that in HSV color space. The precision is about 48–50% in RGB color space when the number of color quantization is 64, while the precision is about 47–49% in HSV color space when the number of color quantization is 72. In other words, the total number of quantization bins in HSV color space is higher than that in RGB color space, but its image retrieval precision is lower than that of RGB color space. Considering that the color quantization level determines the feature vector dimensionality, we select the RGB color space for color quantization in the proposed MTH scheme. However, it should be stressed that this does not mean that RGB color space will also be better than HSV color space in other image retrieval methods. It only validates that RGB is better fitted for the proposed MTH. Indeed, HSV color space is widely use in image retrieval and object recognition and achieves good performance [3,26,38,39]. Based on the results in Table 1 and in order to balance the retrieval precision and vector dimensionality, the final number of color quantization and texture orientation quantization in the proposed MTH are set to 64 and 18, respectively.

To validate the performance of the proposed texture orientation detection method proposed in Section 3.1, we used several typical gradient operators to detect the gradient magnitude and orientation and listed the image retrieval results in Table 3. Note that the proposed method works on the full color image, while the other four operators work on the gray level version of the color images. It can be seen from Table 3 that the proposed orientation

Table 1
The average retrieval precision of MTH with different texture orientation quantization and color quantization levels on the Corel-5000 dataset in RGB color space.

Color quantization levels	Texture orientation quantization levels											
	Precision (%)						Recall (%)					
	6	9	12	18	24	36	6	9	12	18	24	36
128	50.77	51.43	51.22	51.25	51.32	51.14	6.09	6.17	6.15	6.15	6.16	6.13
64	48.82	49.43	49.85	49.98	50.08	49.52	5.86	5.93	5.98	6.00	6.01	5.94
32	45.95	46.93	47.43	47.93	48.00	47.48	5.51	5.63	5.69	5.75	5.76	5.70
16	41.88	42.76	43.42	44.20	44.25	44.43	5.03	5.13	5.21	5.30	5.31	5.33

Table 2
The average retrieval precision of MTH with different texture orientation quantization and color quantization levels on the Corel-5000 dataset in HSV color space.

Color quantization levels	Texture orientation quantization levels											
	Precision (%)						Recall (%)					
	6	9	12	18	24	36	6	9	12	18	24	36
192	48.38	48.77	49.22	49.90	49.78	50.05	5.81	5.85	5.91	5.99	5.97	6.01
128	48.13	48.62	49.07	49.85	49.83	50.38	5.78	5.83	5.89	5.98	5.98	6.05
108	47.95	48.70	49.00	49.37	49.92	49.87	5.75	5.84	5.88	5.92	5.99	5.98
72	47.70	48.15	49.05	49.23	49.41	49.48	5.72	5.78	5.89	5.91	5.93	5.94

detector achieve better results because it exploits the chromatic information that is ignored by other gradient operators in orientation detection.

We then validate the performance of our distance metric and other popular distance or similarity metrics in the proposed MTH method. As can be seen from Table 4, the proposed distance metric obtains much better results than other others distance metrics or similarity metric such as histogram intersection. We can also see that the L_1 distance and L_2 Euclidian distance have the same result with the proposed MTH method, but L_1 distance is much more computationally efficient at the price of losing rotation invariant property [38].

The proposed MTH integrates the merits of co-occurrence matrix and histogram by representing the attribute of co-occurrence matrix using histogram. As can be seen form Fig. 5, there are many bins whose frequencies are close to zero, thus if we apply histogram intersection to MTH. The probability that $\min(T_i, Q_i) = 0$ will be high and hence false match may appear. Therefore, histogram intersection is not suitable to the proposed MTH as a similarity metric. The results in Table 4 also validate this. Meanwhile, the proposed distance metric in Section 4.2 is simple to calculate, while it can be considered as a weighted L_1 distance with the $1/(1+T_i+Q_i)$ being the weight. Since the same values of $|T_i-Q_i|$ can come from different pairs of T_i and Q_i, using the weight parameter can reduce the opposite forces.

We vary the distance parameter $D=1, 2,\ldots,9$ in calculating the MTH. The average retrieval precision values are listed in Table 5. The average retrieval precision of MTH is from about 49–48% for Corel-5000 dataset and from about 40–39% for Corel-10 000 dataset. The best performance of MTH is obtained when $D=1$ for both Corel 5000 dataset and Corel 10 000 dataset. MTH takes into account the spatial correlation between neighboring color and edge orientation by using the four texton types. If we increase the values of distance parameter, the performance is reduced because the probability of neighboring pixels with the same gray level in a 2×2 grid is higher than that in a bigger grid. In other word, the information with $D=1$ is richer than other distance values, thus MTH obtains the best performance when the distance parameter $D=1$.

The average retrieval precision and recall results on the two datasets are listed in Table 6, and the average retrieval precision and recall curves are plotted in Fig. 6. It can be seen from the Table 6 and Fig. 6 that our method achieves much better results than EOAC and TCM methods. On the Corel-5000 dataset with $D=1$, MTH's precision is 22.62% and 18.75% higher than TCM and EOAC, respectively. On the Corel-10 000 MTH's precision is 20.45% and 17.51% higher than TCM and EOAC, respectively.

Figs. 7 and 8 show two retrieval examples on the Corel 5000 and Corel 10 000 datasets. In Fig. 7, the query is a stained glass image, and the top all retrieved images show good match of texture and color to the query image. In Fig. 8, the query image is a racing car which has obvious shape features. All the top 12 retrieved images show good match of the shape, where 10 returned images belong to F1 racing car.

Table 3
The retrieval precision of MTH with different gradient operators for orientation detection.

Dataset	Performance	Gradient operators				
		Proposed	Sobel	Robert	LoG	Prewitt
Corel-5000	Precision (%)	49.98	49.58	48.93	48.02	49.24
	Recall (%)	6.00	5.45	5.87	5.76	5.91
Corel-10 000	Precision (%)	40.87	39.48	39.18	38.72	39.26
	Recall (%)	4.91	4.74	4.71	4.65	4.71

Table 6
The average retrieval precision and recall results on the two Corel datasets.

Datasets	Performance	Method		
		EOAC	TCM	MTH
Corel-5000	Precision (%)	31.23	27.36	49.98
	Recall (%)	3.74	3.28	6.00
Corel-10 000	Precision (%)	23.36	20.42	40.87
	Recall (%)	2.81	2.45	4.91

Table 4
The average retrieval precision of MTH with different distance metrics.

Dataset	Performance	Distance or similarity metrics			
		Our distance metric	L_1	Euclidian	Histogram intersection
Corel-5000	Precision (%)	49.98	45.55	45.55	35.62
	Recall (%)	6.00	5.47	5.47	4.27
Corel-10 000	Precision (%)	40.87	35.29	35.29	27.37
	Recall (%)	4.91	4.23	4.23	3.28

Table 5
The average retrieval precision of MTH with different distance parameter.

Datasets	Performance	Distance parameter (***D***)								
		1	2	3	4	5	6	7	8	9
Corel-5000	Precision (%)	49.98	49.37	49.10	49.30	49.22	49.08	49.07	48.63	48.47
	Recall (%)	6.00	5.93	5.89	5.92	5.91	5.89	5.89	5.84	5.82
Corel-10 000	Precision (%)	40.87	40.79	40.61	40.33	40.26	40.18	40.02	39.86	39.52
	Recall (%)	4.91	4.89	4.87	4.84	4.83	4.82	4.80	4.78	4.74

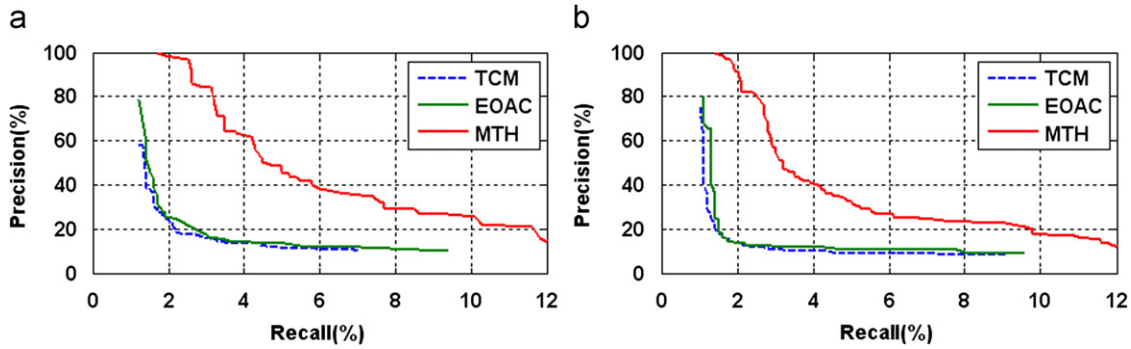

Fig. 6. The precision and recall curves of EOAC, TCM and MTH. (a) Corel-5000 dataset and (b) Corel-10 000 dataset.

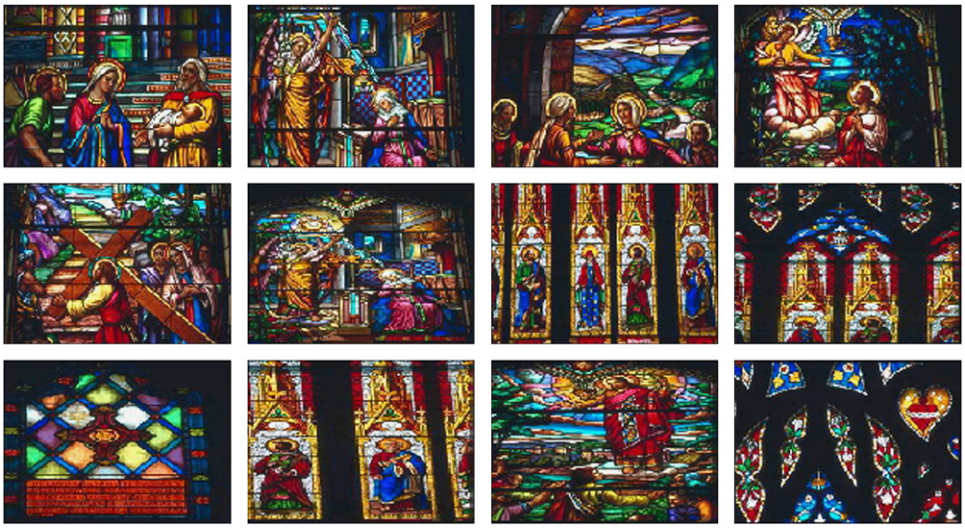

Fig. 7. An example of image retrieval by MTH on the Corel 5000 dataset. The query is a stained glass image, and all images are correctly retrieved and ranked within the top 12 images. (The top-left image is the query image, and the similar images include the query image itself).

Fig. 8. An example of image retrieval by MTH on the Corel 10 000 dataset. The query is a racing car image, and all the returned images are correctly retrieved and ranked within top 12 images, where 10 returned images belong to F1 racing car. (The top-left image is the query image, and the similar images include the query image itself).

EOAC is invariant to translation, scaling, illumination and small rotation. It represents edges features based on their orientations and correlation between neighboring edges. Though EOAC can well represent the shape information of the image, it cannot well represent the color and texture features [27]. In the experiments we see that EOAC achieves good performance only for a few image

categories which have obvious shape features without complex background. EOAC is also appropriate for retrieving images with continuous and clear edges, especially for images with direct lines. However, EOAC is not appropriate for retrieving images with texture and unclear edges [27]. In order to be invariant to illumination, EOAC loses some color information. The spatial correlation of edge and edge orientation can only represent image features partially. EOAC has advantage in shape feature representation by the spatial correlation of edge orientation, and this advantage is preserved in the proposed method.

TCM describes an image by its gradient information and color information with a 12-dimensional vector, including features of energy, contrast, entropy and homogeneity [20]. However, TCM does not take into account the relationship between gradient and color features, and thus the discrimination power of TCM is not high enough for image retrieval in large scale image datasets. The features used in TCM belong to the second-order statistics. Based on Julesz's texton theory, the second-order statistics are not always identical to the difference of two textures [1,2], so using only those features to describe image content may not always enhance the texture discrimination power. The proposed MTH combines the first-order statistics and second-order statistics into an entity for texton analysis, and thus the texture discrimination power is greatly increased. MTH can represent the spatial correlation of edge orientation and color based on textons analysis. So its performance is better than EOAC and TCM.

The experiments were all performed on a double core 1.8 GHz Pentium PC with 1024 MB memory and the Windows XP operating system. The image retrieval system was built in Borland Delphi 7. During the course of features extraction for a natural image of size 192×128, the average time consumption of EOAC, TCM and MTH are 887.40, 157.36 and 314.38 ms, respectively. The time used by MTH is mainly on the stage of texton analysis.

5. Conclusion

We proposed a new method, namely multi-texton histogram (MTH), to describe image features for image retrieval. MTH can represent both the spatial correlation of texture orientation and texture color based on textons. It integrates co-occurrence matrix and histogram into one descriptor and represents the attribute of co-occurrence matrices using histograms. MTH does not need any image segmentation, learning and training stages, and it is very easy to implement. It is well suited for large-scale image dataset retrieval. MTH can be considered as a generalized visual attribute descriptor. Moreover, when used as a color texture descriptor, it can obtain good performance for natural texture extraction. The dimension of MTH feature vector is only 82, which is efficient for image retrieval. The experiments were conducted on two Corel datasets in comparison with the edge orientation auto-correlogram (EOAC) method and the texton co-occurrence matrix (TCM) method. The experimental results validated that our method has strong discrimination power of color, texture and shape features, and outperforms EOAC and TCM significantly.

Acknowledgment

This work was supported by the National Natural Science Fund of China (No. 60632050) and the Hong Kong RGC General Research Fund (PolyU 5351/08E). The authors would like to thank the anonymous reviewers for their constructive comments.

References

[1] B. Julesz, Textons, the elements of texture perception and their interactions, Nature 290 (5802) (1981) 91–97.
[2] B. Julesz, Texton gradients: the texton theory revisited, Biological Cybernetics 54 (1986) 245–251.
[3] R.C. Gonzalez, R.E. Woods, Digital Image Processing, third ed, Prentice Hall, 2007.
[4] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (11) (2007) 262–282.
[5] T. Quack, U. Monich, L. Thiele, B.S. Manjunath, Cortina: a system for large-scale, content-based web image retrieval, in: Proceedings of the 12th annual ACM international conference on Multimedia, 2004.
[6] J.R. Smith, S.-F. Chang, Visual Seek: A Fully Automated Content-Based Image Query System, In ACM Multimedia, Boston, MA, 1996 87–98.
[7] J.Z. Wang, J. Li, G. Wiederholdy, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (23) (2001) 947–963.
[8] F. Monay, D. Gatica-perez, Modeling semantic aspects for cross-media image indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1802–1817.
[9] R. Marée, P. Geurts, L. Wehenkel, Content-based image retrieval by indexing random subwindows with randomized trees, ACCV 4844 (2007) 611–620.
[10] A. Singhal, J. Luo, W. Zhu, Probabilistic spatial context models for scene content understanding, Proceedings of Computer Vision and Pattern Recognition 1 (1) (2003) 1235–1241.
[11] N. Vasconcelos, Image indexing with mixture hierarchies, Proceedings of Computer Vision and Pattern Recognition 1 (1) (2001) 1–10.
[12] X. He, W.-Y. Ma, H.-J. Zhang, Learning an image manifold for retrieval, in: Proceedings of the 12th Annual ACM International Conference on Multimedia, 2004.
[13] S.C.H. Hoi, R. Jin, J. Zhu, M.R. Lyu, Semi-Supervised SVM batch mode active learning and its applications to image retrieval, ACM Transactions on Information Systems (TOIS) 27 (3) (2009) 1–29.
[14] R. Zhang, Z. Zhang, Effective image retrieval based on hidden concept discovery in image database, IEEE Transactions on Image processing 16 (2) (2007) 562–572.
[15] L. Wu, Y. Hu, M. Li, N. Yu, X.-S. Hua, Scale invariant visual language modeling for object categorization, IEEE Transactions on Multimedia 11 (2) (2009) 286–294.
[16] J. Amores, N. Sebe, P. Radeva, Context based object-class recognition and retrieval by generalized correlograms, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1818–1833.
[17] Y. chi, M.K.H. Leung, Part-based object retrieval in cluttered environment, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 890–895.
[18] Y. chi, M.K.H. Leung, ALSBIR: a local-structure-based image retrieval, Pattern Recognition 40 (1) (2007) 244–261.
[19] N. Alajlan, M.S. Kamel, G.H. Freeman, Geometry-based image retrieval in binary image databases, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6) (2008) 1003–11013.
[20] G.-H. Liu, J.-Y. Yang, Image retrieval based on the texton co-occurrence matrix, Pattern Recognition 41 (12) (2008) 3521–3527.
[21] S. Kiranyaz, M. Ferreira, M. Gabbouj, A generic shape/texture descriptor over multiscale edge field: 2-D walking ant histogram, IEEE Transactions on Image processing 17 (3) (2008) 377–390.
[22] C.-H. Yao, S.-Y. Chen, Retrieval of translated, rotated and scaled color textures, Pattern Recognition 36 (4) (2003) 913–929.
[23] T. Gevers, A.W.M. Smeulders, PicToSeek: combining color and shape invariant features for image retrieval, IEEE Transactions on Image processing 9 (1) (2000) 102–119.
[24] A.K. Jain, A. Vailaya, Image retrieval using color and shape, Pattern Recognition 29 (8) (1996) 1233–1244.
[25] A.K. Jain, A. Vailaya, Shape-based retrieval: a case study with trademark image database, Pattern Recognition 31 (9) (1998) 1369–1390.
[26] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, A. Yamada, Color and texture descriptors, IEEE Transactions on Circuit and Systems for Video Technology 11 (6) (2001) 703–715.
[27] F. Mahmoudi, J. Shanbehzadeh, et al., Image retrieval based on shape similarity by edge orientation autocorrelogram, Pattern Recognition 36 (8) (2003) 1725–1736.
[28] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[29] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[30] G. Mori, S. Belongie, J. Malik, Efficient shape matching using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1832–1837.
[31] M. Banerjee, M.K. Kundu, Edge based features for content based image retrieval, Pattern Recognition 36 (11) (2003) 2649–2661.
[32] S. Di Zenzo, A note on the gradient of a multi-image, Computer Vision, Graphics, and Image Processing 33 (1) (1986) 116–125.
[33] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43 (1) (2001) 29–44.

[34] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Proceedings of the 10th IEEE International Conference on Computer Vision, (2005) pp. 1800–1807.
[35] A. Diplaros, T. Gevers, I. Patras, Combining color and shape information for illumination-viewpoint invariant object recognition, IEEE Transactions on Image processing 15 (1) (2006) 1–11.
[36] 〈http://www.ux.uis.no/~tranden/brodatz.html〉.
[37] 〈http://www.outex.oulu.fi/index.php?page=image_database〉.
[38] J. Luo, D. Crandall, Color object detection using spatial-color joint probability functions, IEEE Transactions on Image Processing 15 (6) (2006) 1443–1453.
[39] M.J. Swain, D.H. Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991) 11–32.

About the Author—GUANG-HAI LIU is currently an associate professor with the College of Computer Science and Information Technology, Guangxi Normal University in China. He received Ph.D degree from the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST). His current research interests are in the areas of image processing, pattern recognition and artificial intelligence.

About the Author—LEI ZHANG received the B.S. degree in 1995 from Shenyang Institute of Aeronautical Engineering, Shenyang, PR China, the M.S. and PhD degrees in Electrical and Engineering from Northwestern Polytechnic University, Xi'an, PR China, respectively, in 1998 and 2001. From 2001 to 2002, he was a research associate in the Department of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. Since January 2006, he has been an Assistant Professor in the Department of Computing, The Hong Kong Polytechnic University. His research interests include Image and Video Processing, Biometrics, Pattern Recognition, Multi-sensor Data Fusion and Optimal Estimation Theory, etc.

About the Author—YING-KUN HOU is currently a Ph.D candidate with the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST). He is also a lecturer with the School of Information Science and Technology, Taishan University. His current research interests include image processing, digital watermarking and pattern recognition.

About the Author—ZUO-YONG LI received the B.S. degree in computer science and technology from Fuzhou University in 2002. He got his M.S. degree in computer science and technology from Fuzhou University in 2006. He is a Ph.D. candidate now in Nanjing University of Science and Technology, China. His research interests include image segmentation and pattern recognition.

About the Author—JING-YU YANG received the B.S. Degree in Computer Science from Nanjing University of Science and Technology (NUST), China. From 1982 to 1984 he was a visiting scientist at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1993 to 1994 he was a visiting professor at the Department of Computer Science, Missourian University in 1998; he worked as a visiting professor at Concordia University in Canada. He is currently a professor and Chairman in the department of Computer Science at NUST. He is the author of over 100 scientific papers in computer vision, pattern recognition and artificial intelligence. He has won more than 20 provincial awards and national awards. His current research interests are in the areas of image processing, robot vision, pattern recognition and artificial intelligence.