Pattern Recognition 48 (2015) 2554–2566

Contents lists available at ScienceDirect

Pattern Recognition

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

Content-based image retrieval using computational visual attention model

Guang-Hai Liu [a,*], Jing-Yu Yang [b,*], ZuoYong Li [c]

[a] College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China
[b] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China
[c] Department of Computer Science, Minjiang University, Fuzhou 350108, China

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form
1 December 2014
Accepted 7 February 2015
Available online 16 February 2015

Keywords:
Image retrieval
Gray level co-occurrence matrix
Visual attention
Saliency structure model
Saliency structure histogram

ABSTRACT

It is a very challenging problem to well simulate visual attention mechanisms for content-based image retrieval. In this paper, we propose a novel computational visual attention model, namely saliency structure model, for content-based image retrieval. First, a novel visual cue, namely color volume, with edge information together is introduced to detect saliency regions instead of using the primary visual features (e.g., color, intensity and orientation). Second, the energy feature of the gray-level co-occurrence matrices is used for globally suppressing maps, instead of the local maxima normalization operator in Itti's model. Third, a novel image representation method, namely saliency structure histogram, is proposed to stimulate orientation-selective mechanism for image representation within CBIR framework. We have evaluated the performances of the proposed algorithm on two datasets. The experimental results clearly demonstrate that the proposed algorithm significantly outperforms the standard BOW baseline and micro-structure descriptor.

1. Introduction

With the development of digital image processing technology, large collections of image data have become readily available. Since the demand of market, search or retrieval has become a popular service, where image retrieval has also become a very extensively investigated topic, but how to extract features from the vast amount of image data is a challenging problem. Fortunately, human's visual system has visual attention mechanism that helps humans and primates rapidly select the highly relevant information from a scene. Content-based image retrieval (CBIR) can benefit from visual attention mechanisms by using the saliency information. Recently, developing computational visual-attention models to simulate human visual mechanism has been attracting more and more interest in the field of computer vision. The neural mechanisms of human vision system still do not fully understand, so it is a very challenging problem to build a comprehensive computational model to well simulate visual attention mechanism for image retrieval.

Image retrieval techniques can be widely classified into two categories: (1) the methods based on global features and (2) the methods based on local features. It is worth noting that, whether global features or local features, the extraction of the primary visual features is one of the key issues in image retrieval. Images contain a rich variety of semantic information. Two look very similar images may be different from each other. For instance, in Fig. 1, the leopard and the tiger have similar texture attribute and appearance, but they not belong to the same animal. In some cases, many people cannot distinguish them immediately. Such a problem involves visual attention mechanism. It is well known that the responses of simple cells in V1 to visual stimuli are selective to spatial frequency and orientation [1,2]. Current CBIR technologies are mainly based on global features (e.g., color, texture, edges and spatial information). As mentioned before, CBIR can benefit from visual attention mechanisms, but developing computational visual attention model for image representation within CBIR framework need to be further studied.

In our earlier work [3], micro-structures model is developed for content-based image retrieval. Micro-structures are defined as the collection of certain underlying colors, where the idea of micro-structures model derived from Treisman's feature integration theory [4] and Julesz' texton theory [49,50]. Even so, orientation-selective mechanism and saliency information are not reflected in micro-structures model. In other words, micro-structures model has not simulated visual attention mechanisms well. To address this problem, we propose a novel computational visual-attention model, namely salient structure model, for content-based image retrieval. There are three highlights in this model: (1) a novel visual cue, namely color volume, with edge information together is used to detect saliency regions instead of using the primary visual features (e.g., color, intensity and orientation). (2) the energy feature of gray-level co-occurrence

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

http://dx.doi.org/10.1016/j.patcog.2015.02.005
0031-3203/

Fig. 1. The texton differences between two natural images. The leopard and the tiger have similar texture attribute and appearance, but they not belong to the same animal, where the spatial distribution of their textons is different from each other.

matrices (GLCM) is used for globally suppressing maps, instead of the local maxima normalization operator in Itti's model [5], where the energy feature of GLCM has definite physical meaning and can be considered as a certain prior knowledge. (3) A novel image representation method, namely saliency structure histogram, is proposed to stimulate orientation-selective mechanism for image representation within CBIR framework.

The rest of this paper is organized as follows. In Section 2, we summarize the current knowledge about saliency models and the classical techniques related to image retrieval. In Section 3, gray level co-occurrence matrix is introduced. The proposed saliency model and descriptor are presented in Section 4. In Section 5, performance comparisons among the standard BOW baseline [34], micro-structure descriptor [3] and the proposed algorithm are taken on two datasets. Section 6 concludes the paper.

2. Related works

In this paper, developing a novel computational visual attention model for content-based image retrieval is our scope, where the main concerns are visual attention models and image retrieval techniques. In the following subsections, a comprehensive review of visual attention models is given. Besides, the classical techniques related to image retrieval, feature extraction and image representation are also introduced.

2.1. Visual attention models

Recently, modeling visual attention has raised more and more interest in the field of computer vision. Several visual attention models have been suggested over the past years. Visual attention models can be categorized as bottom-up models and top-down models. Bottom-up models are mainly based on the characteristics of a visual scene and belong to the stimulus-driven model, whereas top-down models are determined by cognition phenomena [13].

Majorities of computational visual attention models derived from the feature integration theory [4] and the Guided search model [6]. Itti's saliency model is the most classical model [5]. It has become the basis of later saliency models and the standard benchmark for comparisons. In Itti's model, an input image is subsampled into a set of Gaussian pyramid for extracting visual features of color, intensity and orientation, respectively. Each feature is computed by a set of center-surround operation akin to visual receptive fields. Finally, various features are combined into a saliency map. Tsotsos et al. have presented a visual attention model based on the concept of selective tuning, where a top-down hierarchy of winner-take-all networks is used for tuning model neurons at the attended locations [7]. Walther and Koch have proposed a biologically plausible model of forming and attending to proto-objects in natural scenes [8]. Meur et al. have proposed a coherent computational approach to the modeling of the bottom-up visual attention based on the physical structure of human visual system [9]. Sun and Fisher have developed a novel model of object-based visual attention according to two new mechanisms. The first mechanism computes the visual salience of objects and groupings; the second one implements the hierarchical selectivity of attention shifts [10]. Borji and Itti have introduced a model for exploiting local and global patch rarities to detect saliency areas [11]. In addition to the above saliency models, there are many other saliency models which are based on probabilistic, psychophysics or neurophysiology models. A comprehensive review of visual attention models can be found in [12,13] and will not be considered here.

In summary, saliency models have been well studied in the field of computer vision, but there are few published articles to investigate visual attention model within the CBIR framework. Moreover, how to construct visual attention model is still an open problem.

2.2. The classical techniques related to image retrieval

The classic image retrieval techniques are based on two types of visual features: global features and local features. Global features-based algorithms aim at the whole image as visual content, e.g. color, texture and shape, and local features-based algorithms focus mainly on keypoints or salient patches. Various algorithms have been designed for extracting global and local features.

In the MPEG-7 standard, the color descriptors consist of a number of histogram descriptors, such as dominant color descriptor, color layout descriptor, and scalable color descriptor [15]. Various algorithms have been developed for texture analysis in some literatures, such as the gray level co-occurrence matrices [16], Tamura texture feature [17], Markov random field model [18], Gabor filtering [19], local binary pattern [20], etc. There are three texture descriptors in the MPEG-7 standard: homogeneous texture descriptor, texture browsing descriptor and the edge histogram descriptor [15]. Texture features can be combined with color features to improve the discrimination power and can obtain better retrieval performance. There are some algorithms which can ultimately combine color and texture together, such as texton co-occurrences matrix [21], multi-texton histogram [22], and color edge co-occurrence histogram [23], micro-structure descriptor [3], color difference histogram[24] etc. The classical representations of shape feature include moment invariants [25,26], Fourier transforms coefficients [27,28], edge curvature and arc length [44]. In MPEG-7 standard, three shape descriptors are used for object-based image retrieval: 3-D shape descriptor, region-based shape descriptor, and curvature scale space (CSS) descriptor [15]. In many cases, shape feature extraction need image segmentation which is still an open problem, and limited its application in many fields, thus many researchers have adopted local features (e.g., keypoints, salient patches) instead of using the traditional shape features.

There are many famous keypoints detectors and descriptors [29–33] [59,60], such as Harris keypoints detector, SIFT, SURF, PCA-SIFT and ORB (oriented FAST and Rotated BRIEF), where SIFT is the most popular local feature representation. It can be used to perform reliable matching between different views of an object or scene [29]. In order to perform as good as SIFT with lower computational complexity, the SURF [32] or ORB [59] can be considered as an efficient alternative to SIFT. Recently, bag-of-visual words (BOW) models or its variants have been reported in the literatures and used for object-based image retrieval, object recognition and scene categorization [34–41]. In [34], Sivic and Zisserman have proposed the bag-of-visual words (BOW) model which in essence borrows techniques from text retrieval. In BOW model, local features extracted from an image by using SIFT, SURF or other keypoints detectors, and then mapped into a set of visual words. Finally, an image is represented as a histogram of visual word occurrences. It is so called the standard BOW baseline, and can be considered as one of state-of-the-art methods. Since the visual words usually come from clustering implementation which needs heavy computational burdens. Besides, visual words have two major limitations that the lack of any explicit semantic meanings and the ambiguity of visual words. Indeed, improving the visual vocabulary, incorporating spatial information and semantic attributes can reduce the limitations and can also improve the performances of BOW models [35–41].

There are extensive studies in feature extraction and image representation within image retrieval and object recognition framework. However, developing computational visual-attention model within CBIR framework needs to be further studied.

3. Gray level co-occurrence matrix (GLCM)

Before discussing the proposed computational visual-attention model in more details, a brief introduction of gray level co-occurrence matrix (GLCM) is given, since our saliency model involves Haralick's gray level co-occurrence matrix [16].

Co-occurrence matrix is the most famous statistical approach in textural image processing. In 1973, Haralick have put forward the gray level co-occurrence matrix, and extracted a set of 14 features to describe texture images features, such as energy, inverse difference moment, contrast, entropy and so on [16]. It remains popular today by virtue of good performance. The value of a gray image at any coordinates(x, y) is denoted as $f(x, y) = w$, $w \in \{0, 1, \ldots, 255\}$. In order to conveniently define the co-occurrence matrix, the pixel position at the coordinates(x, y) is denoted as P, where $P=(x, y)$. Let there are two pixel positions $P_1=(x_1, y_1)$ and $P_2=(x_2, y_2)$, their pixel values are $f(P_1)=w$ 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, the cell entry $(w, \hat{w})$ of co-occurrence matrix $GLCM(w, \hat{w}, d)$ can be defined as follows:

$$GLCM(w, \hat{w}, d) = pr\{f(p_1) = w \wedge f(p_2) = \hat{w} \mid |p_1 - p_2| = d\} \quad (1)$$

where $\wedge$ denotes the logical AND operation. In GLCM algorithm, energy, entropy, contrast and inverse difference moment often utilized to describe image features [16], but the discrimination power does not enough to achieve the satisfactory performance of image retrieval especially on larger scale datasets [21]. If all cell entries of co-occurrence matrix are used to describe image features, the vector dimension would be very high and is not always increase retrieval accuracy.

However, some features extracted from GLCM have definite physical meaning in texture image analysis, where energy is a measure of textural uniformity of an image. When the image under consideration is homogenous, energy reaches its maximum [43]. The conspicuity areas can be considered as those areas which have significant visual differences and are not the homogenous areas. Inspired by above views, the energy feature of GLCM is used as the inhibition term in the stage of saliency map detection, instead of using the local maxima normalization operator in Itti's model [5].

4. The saliency structures model and descriptor

Human's visual attention consists of pre-attentive and attentive stage according to Treisman's feature integration theory [4]. In the pre-attentive stage, only “pop-out” features are detected. Whereas in the attentive stage, relationships between various features are found and grouping [4,14]. In this paper, saliency structure model is proposed to content-based image retrieval according to Treisman's feature integration theory [4] and Julesz' texton theory [49,50]. In feature extraction and image representation, Orientation-selective mechanism which derived from the works of Hubel and Wiesel is used to our model [1]. Color, intensity and orientation are considered as the primary visual features which are commonly used in many saliency models [4,5]. In order to detect “pop-out” features, a novel visual cue, namely color volume, with edge information together is introduced into our saliency model and used to detect saliency regions.

It is crucially important to emphasize that saliency structure model can be considered as an improved version of micro-structures model by combining a bottom-up component of visual attention and orientation-selective mechanism, where the saliency structures are defined as the bar-shaped structures according to orientation-selective mechanism by using oriented Gabor filters, whereas micro-structures are defined as the collection of certain underlying colors [3]. The basic principle of the proposed descriptor is to generate three tuples histograms considering the bar-shaped structures and oriented Gabor filters via a very special type, whereas micro-structure descriptor is adopted the probability statistics method to describe features.

The flow diagram of the proposed saliency model within CBIR framework is illustrated in Fig. 2.

In the proposed saliency model within CBIR framework, we mainly focus on: (1) the construction of saliency structure model and (2) image representation. Where the construction of saliency structure model mainly consists of three stages: (a) extraction of the primary visual features, (b) the saliency map detection and (c) the combination of bar-shaped structure and oriented Gabor filters for saliency structure detection.

4.1. Extraction of the primary visual features

Human's visual system is more sensitive to color, orientation and intensity information [5]. In many visual saliency models, color is implemented as R-G (red-green) and B-Y (blue-yellow) channels inspired by color-opponent neurons in V1 cortex [5] [13]. The average of three color channels is usually used as intensity. Orientation is often implemented as a convolution with oriented Gabor filters.

It is well known that HSV color space could mimic human's color perception well. In order to extract the primary visual features for image representation and simplify manipulation, the quantization of visual features needed to be implemented in HSV color space. For example, the task of color quantization is to select and assign a limited set of colors for representing a give color image with maximum fidelity [44]. The color quantization techniques are more fully described in many books of digital images processing and will not be described in detail here.

In order to obtain color map, H, S and V color channels are uniform quantized into 6, 3 and 3 bins, respectively, so that in total $6 \times 3 \times 3 = 54$ color combinations are obtained, $M_C(x, y)$ denotes the color combinations or color map, as $M_C(x, y) = w$, $w \in \{0, 1, \ldots, N_C - 1\}$, where $N_C = 54$ in this paper.

Intensity information is given by V color channel. After uniform quantization, we can obtain the intensity map $M_I(x, y)$, as $M_I(x, y) = s$, $s \in \{0, 1, \ldots, N_I - 1\}$, where $N_I = 16$. Since the computational

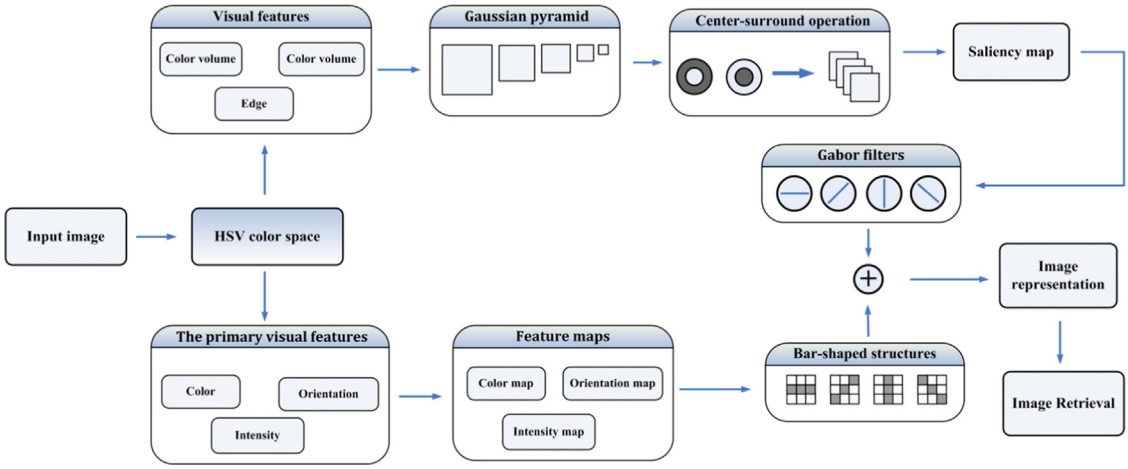

Fig. 2. Flow diagram of the proposed saliency model within CBIR framework.

burden of Gabor filters is much high, so Gabor filters are not used to detect local orientation information in the extraction of primary visual features.

In this paper, intensity information is also used to detect edge orientation map $O(x,y)$ and gradient image $g(x,y)$ by using Sobel operator. After uniform quantization, we can obtain the edge orientation map $M_O(x,y)$, as $M_O(x,y)=\theta,\ \ \theta\in\{0,1,\ldots,N_O-1\}$, where $N_O=60$.

In this paper,$M_C(x,y)$,$M_I(x,y)$ and $M_O(x,y)$ play an important role on saliency structures detection and image representation. More details can be referred to Sections 4.4 and 4.5.

4.2. The saliency map

Most of the existing saliency model frameworks often use a combination of the primary visual features to generate saliency maps, such as intensity, orientation and color information, and so on. Indeed, it is not all the primary visual features can obtain good results in saliency detection. The choice of visual features for saliency detection depends on the real application. Because various visual features provide different contribution to saliency detection, a certain feature may be strong in one case but weak in another. Inspired by this view, we propose a new visual feature, namely color volume, with edge information $g(x,y)$ together to detect saliency areas.

Indeed, both HSV and Lab color space could mimic human's color perception well. The shape of CIE chromaticity diagram looks like a horseshoe, so it is very difficult to calculate the color volume of Lab color space. It is one of the reasons why HSV color space is used to compute color volume. The shape of HSV color space can be interpreted as cylinder coordinate system. It is well known that the cylinder volume cv can be defined as $cv=\pi r^2 h'$, where r denotes the radius of cylinder, and h' denotes the height of cylinder. Let there is a random dot (h,s,v) in cylinder coordinate system, the cylinder volume derived from this dot can be defined as

$$cv_1(x,y)=\pi\times s(x,y)^2\times v(x,y)\times\frac{h(x,y)}{360} \tag{2}$$

where $s(x,y)\in[0,1]$, $v(x,y)\in[0,1]$ and $h(x,y)\in[0,360]$. When HSV color space to be transformed into Cartesian coordinate system, the color volume derived from this dot (h,s,v) can be defined as

$$cv_2(x,y)=s(x,y)\times\ \cos(h(x,y))\times s(x,y)\times\ \sin(h(x,y))\times v(x,y) \tag{3}$$

After the definition of color volume, $cv=\{cv_1,cv_2\}$ and $g(x,y)$ are used to create a Gaussian pyramid $cv(\sigma)$ and $g(\sigma)$, where $\sigma=[0,1,\ldots,4]$ is the scale. As can be seen from Fig. 3(a)–(c), the color conspicuity areas can be straightforward pop-out by using the color volume information $cv=\{cv_1,cv_2\}$, and suppressed the background to certain extent. It is also one of the reasons why HSV color space is used to compute color volume. The standard deviation of Gaussian kernel is $\delta=5.0$. Center-surround receptive fields are simulated by across scale subtraction "Θ" between two maps at center scale(c) and surround scale(s) in pyramids $cv(\sigma)$ and $g(\sigma)$, and yield the so-called feature maps:

$$F(c,s,cv)=|\,cv(c)\Theta cv(s)|\quad\ \forall cv\in\{cv_1,cv_2\} \tag{4}$$

$$F(c,s,g)=|\,g(c)\Theta g(s)| \tag{5}$$

After center-surround operation, we can obtain 18 feature maps. In order to simulate the local competition between neighboring salient image locations, a normalize weight function is defined as follow:

$$w_d(d_1,d_2)=|\,GE(d_2)-GE(d_1)| \tag{6}$$

where $GE(d_i)$ denote the energy feature of the gray-level co-occurrence matrices by using different parameter d_i, $i=1,2$, and d_i denote the offsets between the two pixels. In this paper, we set $d_1=3$ and $d_2=9$ to calculate the energy feature of GLCM. In the field of texture analysis, energy is a measure of textural uniformity of an image. As mentioned before, when the image under consideration is homogenous, energy reaches its maximum [43]. In such a case, there are no significant differences in the image, and the homogenous areas should be suppressed.

Feature maps are combined into three conspicuity maps. In this paper, we denote $\overline{cv}$ and $\overline{g}$ as the individual saliency maps at the scale ($\sigma=4$), where $\overline{cv}$ for color volume, and $\overline{g}$ for edge information. They are obtained through across-scale addition "$\oplus$", which consist of reduction of each map to the scale ($\sigma=4$) and point-by-point addition. It is similar to the manner of Itti's saliency model [5].

$$\overline{cv}=\sum\nolimits_{cv\in\{cv_1,cv_2\}}\{\oplus_{c=0}^{2}\ \oplus_{s=3}^{4}|\,w_d\times F(c,s,cv)|\,\} \tag{7}$$

$$\overline{g}=\ \oplus_{c=0}^{2}\ \oplus_{s=3}^{4}|\,w_d\times F(c,s,g)| \tag{8}$$

The individual saliency maps finally combined into a single overall saliency map, it is calculated as follows:

$$MS=\tfrac{1}{3}(\overline{cv_1}+\overline{cv_2}+\overline{g}) \tag{9}$$

At last, the single overall saliency map MS would be resized until it has the same size as the original image. As can be seen from Fig. 3(e), using our algorithm to detect saliency areas, the partial background of image is shielded and can pop out the major objects.

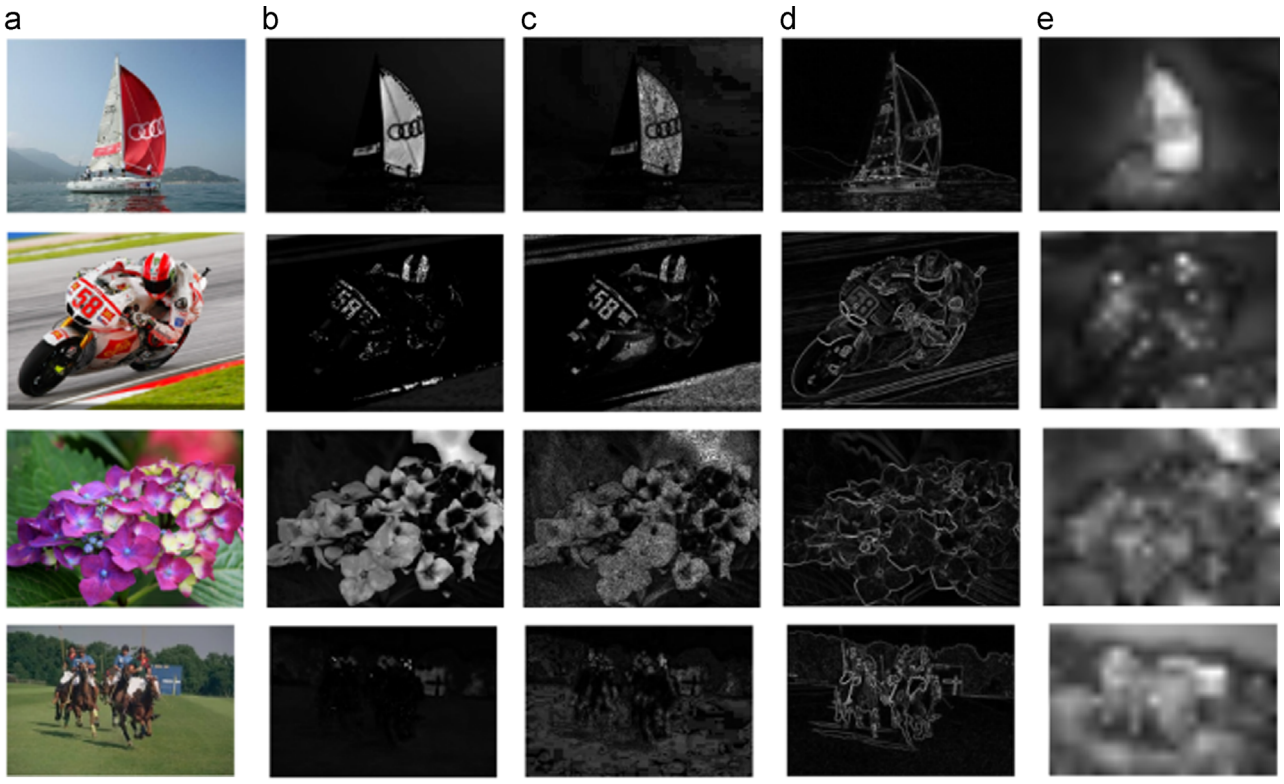

Fig. 3. The illustration examples of the visual features are used to detect saliency maps: (a) original image, (b) color volume $v_1(x,y)$, (c) color volume $v_2(x,y)$, (d) gradient image $g(x,y)$ and (e) saliency map.

After saliency map detection, the single overall saliency map *MS* is used for further processing, and then oriented Gabor filters are introduced to detect saliency structures and describe image features.

4.3. Oriented Gabor filters

It is well known that the receptive field of a simple cell in primary visual cortex (V1) can be accurately modeled by two-dimensional Gabor function, and Gabor energy can capture typically fundamental characteristics of complex cells. Two-dimensional Gabor filter can be defined as [12,46,47]

$$\begin{cases} g(x,y,\varphi,\theta) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \\ x' = x\ \cos\ \theta + y\ \sin\ \theta \\ y' = -x\ \sin\ \theta + y\ \cos\ \theta \end{cases} \tag{10}$$

where θ is the preferred orientation, as $\theta \in [0,\pi)$,γ is the spatial aspect ratio that determines the eccentricity of Gaussian envelope. λ is the wavelength,δ is the standard deviation of Gaussian factor determines the size of receptive filed, andφ is a phase offset and determining the symmetric of $g(x,y,\varphi,\theta)$[12,47]. Gabor energy $E(x,y,\theta)$ can be defined as

$$\begin{cases} E(x,y,\theta) = \sqrt{e(x,y,0)^2 + e\left(x,y,-\frac{\pi}{2}\right)^2} \\ e(x,y,\varphi) = MS(x,y) \otimes g(x,y,\varphi,\theta) \end{cases} \tag{11}$$

where $MS(x,y)$ is the single overall saliency map. Gabor energy $E(x,y,\theta)$ for a number of N_θ can be further defined as $E(x,y,\theta_i)$, where different orientations θ_i can be computed as

$$\theta_i = \frac{(i-1)\pi}{N_\theta}, \quad i = 1,2,\ldots,N_\theta \tag{12}$$

In this paper, $N_\theta = 4$, $\gamma = 0.50$, $\lambda = 7.0$, $\delta = 2.333$ are used to compute Gabor energy. In order to reduce the computational burden, Gabor filters are truncated to 9×9 pixels, where the results of Gabor filtering are not normalized. Too much orientation only slight improves the effectiveness of algorithm whereas computational burden increases, thus only four oriented Gabor energy maps $E(x,y,\theta_i)$, $i = \{1,2,3,4\}$ are used to detect saliency structures and describe image features.

4.4. Saliency structure detection

In real-world, orientation is a powerful visual cue about the subject depicted in an image. Strong orientations usually indicate a definite pattern. However, natural scenes usually do not show strong orientation and the subject has no clear structure [48]. Although the natural images show various contents, they may have some commonly basic elements. The different combination and spatial distribution of those basic elements can result in various local structures or patterns in the natural images. In [3,21,22], several texton patterns are proposed for image analysis. Those texton patterns have their advantages in image content analysis, but orientation-selective mechanism and saliency information are not reflected. Moreover, texton patterns are only used as a tool to detect out the areas sharing a common property all over the image. This manner has limited their advantages and cannot simulate human's visual attention well.

In our earlier work, micro-structures are defined as the collection of certain underlying colors inspired by Julesz's texton theory [49,50]. Besides, micro-structures can be considered as the extension of Julesz's textons or the color version of textons [23]. Indeed, Julesz' texton theory focus on the study of pre-attentive texture discrimination [49,50], and texton also has close relationship with the primary visual cortex (V1). However, orientation-selective mechanism and saliency information are not reflected in micro-structures model.

It is well known that the receptive fields of simple cell in V1 have orientation selection [1,51,52]. In [1], Hubel and Wiesel described simple cells as linear with bar-shaped or edge-shaped receptive fields,

it led to a view of the cortex as containing a population of feature detectors tuned to edges and bars of various widths and orientations [1,53]. This view also led to the proposed saliency structures. Accordingly, our proposed saliency structures derive from orientation-selective mechanism.

In order to simulate orientation-selective mechanism well, we define the bar-shaped structures for image content analysis. It is crucially important to emphasize that bar-shaped structures are quite different from the edge, corner and orientation detection methods [42]. In an image, edge pixels are pixels at which the intensity changes abruptly, and edges (or edges segments) are sets of connected edge pixels [45]. A corner is defined as a location that exhibits a strong gradient value in multiple directions at the same time [44]. The bar-shaped structures are defined as three consecutive adjacent pixels which have the same pixel values. The working mechanisms of bar-shaped structure detection can be introduced as follows:

Let there is a 3×3 block in intensity map $M_I(x,y)$, where (x,y) is discrete coordinate. The value of the center coordinate (x_0,y_0) at the block is denoted as $M_I(x_0,y_0)$. Let there are two coordinates (x_1,y_1) and (x_2,y_2) on both sides of the central coordinate (x_0,y_0), respectively. If $M_I(x_1,y_1)=M_I(x_0,y_0)=M_I(x_2,y_2)$, such a structure can be considered as a bar-shaped structure of intensity map $M_I(x,y)$. In this case, the angle between the bar-shaped structure and horizontal direction is denoted as $\alpha, \alpha=\{0^\circ,45^\circ,90^\circ,135^\circ\}$. The bar-shaped structures of color map $M_C(x,y)$ and edge orientation map $M_O(x,y)$ are computed by using the same computational steps. An example of the bar-shaped structure detection in intensity map $M_I(x,y)$ is shown in Fig. 4.

After the bar-shaped structure detection, four bar-shaped structures can be obtained, and those bar-shaped structures have significant direction-sense about 0°, 45°, 90° and 135°, respectively. It is shown in Fig. 5(a). When a bar-shaped structure meet oriented Gabor filters $g(x,y,\varphi,\theta),\theta=\{0^\circ,45^\circ,90^\circ,135^\circ\}$ which are shown in Fig. 5(b). If $\alpha=\theta$, such a bar-shaped structure can be considered as a saliency structure. The process of saliency structure detection in intensity map $M_I(x,y)$ is shown in Fig. 5(c), where the preferred orientation of oriented Gabor filters is $\theta=45^\circ$. As can be seen from Fig. 5(d), four saliency structure patterns can be obtained, the angle α between the bar-shaped structures and horizontal direction are 0°, 45°, 90° and 135°, respectively. Their corresponding Gabor energy are defined as $ES(x,y,I,\theta)$ in this paper, where $ES(x,y,I,\theta)=E(x,y,\theta),\theta=\{0^\circ,45^\circ,90^\circ,135^\circ\}$.

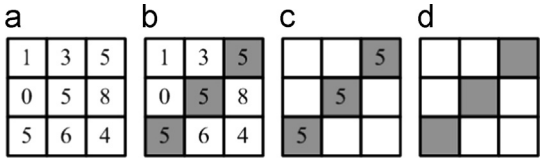

Fig. 4. An example of bar-shaped structure detection in intensity map $M_I(x,y)$, (a) a 3×3 grid of intensity map, (b) and (c) show the process of bar-shaped structure detection and (d) shows the detected bar-shaped structure.

The saliency structure patterns of color map $M_C(x,y)$ and edge orientation map $M_O(x,y)$ are computed by using the same computational steps, and their corresponding Gabor energy are defined as $ES(x,y,C,\theta)$ and $ES(x,y,O,\theta)$, respectively, where $ES(x,y,C,\theta)=E(x,y,\theta)$ and $ES(x,y,O,\theta)=E(x,y,\theta)$, $\theta=\{0^\circ,45^\circ,90^\circ,135^\circ\}$.

It should be stressed that saliency structure patterns govern our perception of the natural images according to orientation-selective mechanism. It can reflect significant direction-sense about 0°, 45°, 90° and 135°, compared to TCM [21], MTH [22] and micro-structures model [3], whereas the textons using in TCM, MTH and micro-structures model have not such a property.

In the field of texture analysis, texture has the spatial repeating properties, similar vectors of responses will reoccur as texture features reoccur in the images [54]. In real-world, the natural images show various contents, only partial belong to texture. However, various images have different color, intensity and edge orientation information. There are some commonly fundamental elements which have the same or similar direction-sense about color, edge orientation and intensity in a natural image or scene. In this paper, the commonly fundamental elements are considered as the saliency structures. The saliency structures derivate from different primary visual features can represent various content types of an image.

We consider saliency structures have spatial repeating properties, similar vectors of responses will reoccur as saliency structure features reoccur in the images. It is the most critical idea of saliency structure histogram or image representation in Section 4.5.

4.5. Feature representation

The central problem in image retrieval or objects recognition tasks can be regarded as extracting "meaningful" features from images [14]. In other words, image representation is the most critical problem of image retrieval or objects recognition. As mentioned before, human's visual system is more sensitive to the primary visual attributes (e.g. color, orientation and intensity information) and has orientation-selective mechanism [5]. Besides, spatial information plays an important role in feature representation. In histogram-based image representation, embedding spatial information can significantly improve the discrimination power of histogram, especially for color histogram. In practice, how to integrate the primary visual features and spatial information into image representation in the manner of simulating orientation-selective mechanism is a challenging problem. To address

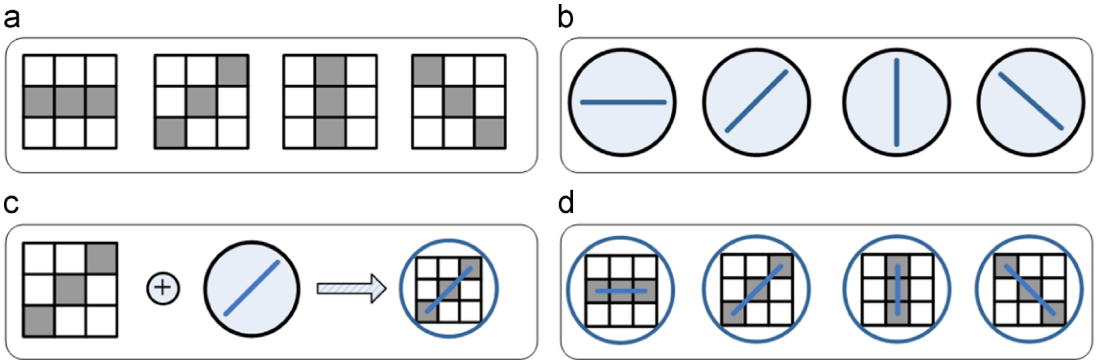

Fig. 5. Saliency structure detection: (a) four bar-shaped structure patterns, the angle α between the bar-shaped structure and horizontal direction are 0°, 45°, 90° and 135°, respectively; (b) four oriented Gabor filters, where the preferred orientation θ are 0°, 45°, 90° and 135°, respectively and (c) the process of saliency structure detection, where term "$\oplus$" denotes that bar-shaped structure meet the oriented Gabor filters. If $\alpha=\theta$, such a bar-shaped structure can be considered as a saliency structure and (d) four saliency structure patterns.

this problem, a novel image feature representation method, namely saliency structure histogram (SSH), is proposed for content-based image retrieval.

In the proposed saliency model which is shown in Fig. 2, color map $M_C(x,y)$, intensity map $M_I(x,y)$ and edge orientation map $M_O(x,y)$ can be obtained after the extraction of the primary visual features, where (x,y) is discrete coordinate and $0 \le x \le wid-1$, $0 \le y \le hei-1$, wid denotes the width of the image and hei denotes the height of the image. The respective quantization number of $M_C(x,y)$, $M_I(x,y)$ and $M_O(x,y)$ are N_C, N_I and N_O. Let there is a 3×3 block in a full color image, where (x,y) denotes the central coordinate of the 3×3 block, moving the 3×3 block from left-to-right and top-to-bottom throughout the full color image with one pixel as the interval. If the saliency structures occur in the 3×3 block, and their corresponding Gabor energy are denoted as $ES(x,y,C,\theta)$,$ES(x,y,O,\theta)$ and $ES(x,y,I,\theta)$, where $\theta=\{0^\circ,45^\circ,90^\circ,135^\circ\}$, then the saliency structure histogram (SSH) can be defined as follows:

$$SSH = conca\{[H_C]^+,[H_O]^+,[H_I]^+\} \tag{13}$$

where

$$H_C(M_C(x,y)) = \begin{cases} \log\left\{\sum_{x=0}^{wid-1}\sum_{y=0}^{hei-1} ES(x,y,C,\theta)\right\}, & \theta=\{0^\circ,45^\circ,90^\circ,135^\circ\} \\ 0 \le M_C(x,y) \le N_C-1 \end{cases} \tag{14}$$

$$H_O(M_O(x,y)) = \begin{cases} \log\left\{\sum_{x=0}^{wid-1}\sum_{y=0}^{hei-1} ES(x,y,O,\theta)\right\}, & \theta=\{0^\circ,45^\circ,90^\circ,135^\circ\} \\ 0 \le M_O(x,y) \le N_O-1 \end{cases} \tag{15}$$

$$H_I(M_I(x,y)) = \begin{cases} \log\left\{\sum_{x=0}^{wid-1}\sum_{y=0}^{hei-1} ES(x,y,I,\theta)\right\}, & \theta=\{0^\circ,45^\circ,90^\circ,135^\circ\} \\ 0 \le M_I(x,y) \le N_I-1 \end{cases} \tag{16}$$

where $conca\{.\}$ denotes the concatenation of H_C,H_O and H_I, $[w]^+ = \max(w,0)$ denotes a half-wave rectification operation and setting all $w<0$ values to zero.

Subjective brightness perceived by human's visual system is a logarithmic function of the light intensity incident on the eye [45,57,58]. However, the visual system cannot operate over such a range simultaneously. It accomplishes this large variation by changes in its overall sensitivity [45]. Inspired by this view, the Logarithm characteristic of Gabor energy is used for image feature representation. Indeed, saliency structure histogram can be considered as a special histogram representation, where the Logarithm characteristic of Gabor energy is used as the values of histograms.

In saliency structure histogram, all three cues (i.e., color, edge orientation and intensity) have been incorporated in the unified framework, where H_C can represent the spatial repeating properties of color structures, leading to a 54 dimensional vector; H_O can represent the spatial repeating properties of edge orientation structures, leading to a 60 dimensional vector; H_I can represent the spatial repeating properties of the intensity structures, leading to a 16 dimensional vector, so that result in total $54+60+16=130$ dimensional vector as the final image features in content-based image retrieval.

In saliency structures model, oriented Gabor filters are embedded into bar-shaped structure for saliency structures detection. In this manner, orientation-selective mechanism can be reflected in our model. Besides, the primary visual features (e.g. intensity, color and orientation), spatial layout, saliency information and Gabor energy are integrated into one whole unit.

5. Experiments

In this section, the performances of the proposed algorithm are evaluated on two datasets. In experiments, two subsets of 2000 and 3000 images, randomly chosen from two datasets, were used as query images, and the system performs the similarity evaluation with respect to each query image. The final performances are evaluated by the average results of all queries. The standard BOW baseline can be considered as one of state-of-the-art methods in the field of object-based image retrieval and object recognition. Micro-structure descriptor (MSD) is our earlier work, thus we have selected the standard BOW baseline [34] and micro-structure descriptor (MSD) [3] for comparisons. The online image retrieval system by the proposed algorithm available at: http://www.ci.gxnu.cn/cbir/.

In the standard BOW baseline, SIFT algorithm is used to extract local features, and then local features are mapped into a set of visual words. Finally, an image is represented as a histogram of visual word occurrences. In experiments, the vocabularies of BOW are generated by using the standard K-means clustering, where $k=1000$ and cosine distance is used as the distance metric.

5.1. Datasets

Two datasets are used in our image retrieval systems. The total number of images on the two datasets is 20,000. All images on the two datasets are full color images.

The first one is Corel-10K dataset. Corel images representing various subjects and scenes like flowers, animals, landscapes, people and so on. Thus, Corel dataset is the most commonly used dataset to test content-based image retrieval performance. Corel-10K dataset contains 100 categories and 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 or 128×192 in JPEG format. All Corel images come from Corel Gallery Magic 20, 0000 (8 cds).

The second dataset is GHIM-10K dataset. All images of this dataset are come from web and camera, and collected by the author (*Guang-Hai Liu*). It contains 20 categories. There are 10,000 images from diverse contents such as sunset, tiger, flower, building, car, mountains, fish, etc. Every category contains 500 images of size 400×300 or 300×400 in JPEG format. Most of those original images are high resolution images and be zoomed into size 400×300 or 300×400.

5.2. Distance metric

It is well known that the retrieval accuracy not only depends on strong features representation, but also on good similarity measure or distance metric. How to measure the similarity between two images is still a largely unanswered question. For each template image in the dataset, a K-dimensional feature vector $T=[T_1,T_2,\ldots,T_K]$ will be extracted and stored in the SQL server database. Let $Q=[Q_1,Q_2,\ldots,Q_K]$ is the feature vector of a query image, and then $L1$ distance metric between them is simply calculated as

$$D(T,Q) = \sum_{i=1}^{K} |T_i - Q_i| \tag{17}$$

$L1$ distance is the most simple distance and more suitable for large scale datasets. For the proposed saliency structure histogram, $K=54+60+16=130$ for color images. If $D(T,Q)$ is small, the two images are similar in their image content.

5.3. Performance metrics

In the field of information retrieval, two primary metrics are *precision* and *recall*. The two metrics are often combined as the

weighted harmonic mean, namely $F-measure$, and it is an overall performance measure [55,56]. It can be defined as

$$\begin{cases} F=\frac{(1+\beta^2)\times R\times P}{(\beta^2\times P)+R} \\ P=\frac{I_N}{N} \\ R=\frac{I_N}{M} \end{cases} \tag{18}$$

In the experiments of image retrieval, *precision*(P) is the ratio of the number of retrieved similar images to the number of retrieved images, while *recall*(R) is the ratio of the number of retrieved similar images to the total number of similar images [3,21,22,24]. Where I_N is the number of retrieved similar images, N is the total number of images retrieved and M is the total number of similar images on dataset. Parameter β allows one to weight either *precision* or *recall* more heavily, and they are balanced when $\beta=1$. If there is no particular reason to favor precision or recall, $\beta=1$ is commonly used to image retrieval or information retrieval [56].

In our image retrieval system, we set $N=12$, $M=100$ and $\beta=1$ on Corel-10K dataset, $N=12$, $M=500$ and $\beta=1$ on GHIM-10K dataset. Thus, $F-measure$ is so called $F_1-measure$. All measures are averaged all the queries on the datasets to obtain overall performance data.

5.4. Retrieval performance and discussion

In this section, we first evaluate the retrieval performances in HSV color space and confirm the respective quantization number of color, edge orientation and intensity information. Second, we will test the mask size of Gabor filter on the influence of retrieval performances. Third, the performances of different primary visual features are evaluated. Fourth, we will confirm the contribution of color-volume and bar-shaped structures in the improvement of retrieval performances. Finally, comparisons and retrieval performances are discussed.

5.4.1. Quantization levels investigation

It is very important to select proper vector dimension which can obtain good retrieval performances while not requiring a great amount of storage space and computational burdens. The quantization number of color, orientation and intensity information can clearly influence the retrieval results.

In the experiments, *bin*(.) is denoted as the quantization level of an image component, where we let $bin(H)\geq 6$, $bin(S)\geq 3$, and $bin(V)\geq 3$ in HSV color space, and hence the total number of bins is at least $6\times 3\times 3=54$, and it is gradually increased to $12\times 3\times 3=108$ bins. Besides, different quantization level of edge orientation and intensity are also used to test the performance of saliency structure histogram. The quantization level of edge orientation varied from 18 to 90 bins, and the quantization level of intensity varied from 16 to 64 bins. In human's visual system, there is a phenomenon known as brightness adaptation [45]. Generally speaking, brightness adaptation level of human's eyes is about 64 levels [45,57]. It is the reason why the maximum quantized level of intensity is 64.

As can be seen from Fig. 6, the quantization level of intensity and color are fixed at a certain bins, increase the quantization level of edge orientation, the performance also increases. When the quantization level of edge orientation beyond certain bins (e.g. bin (ori)=60), the performance reduces. In real-world, orientation is a powerful visual cue about the subject depicted in an image. Strong orientation usually indicates a definite pattern. However, fine orientation quantization may be caused noise structures, and coarse orientation quantization may reduce the representation ability of the proposed algorithm. Both of two cases are unsuitable for image representation.

Indeed, all three cues (color, edge orientation and intensity) have influence on the retrieval performances. Besides, it is worth mentioning that Gabor energy of the saliency structures is also control the performances of the proposed algorithm. Accordingly, retrieval performances are the results of four cues (color, edge orientation, intensity, Gabor energy of the saliency structures) work together. Single cue cannot guarantee that the performance always improved.

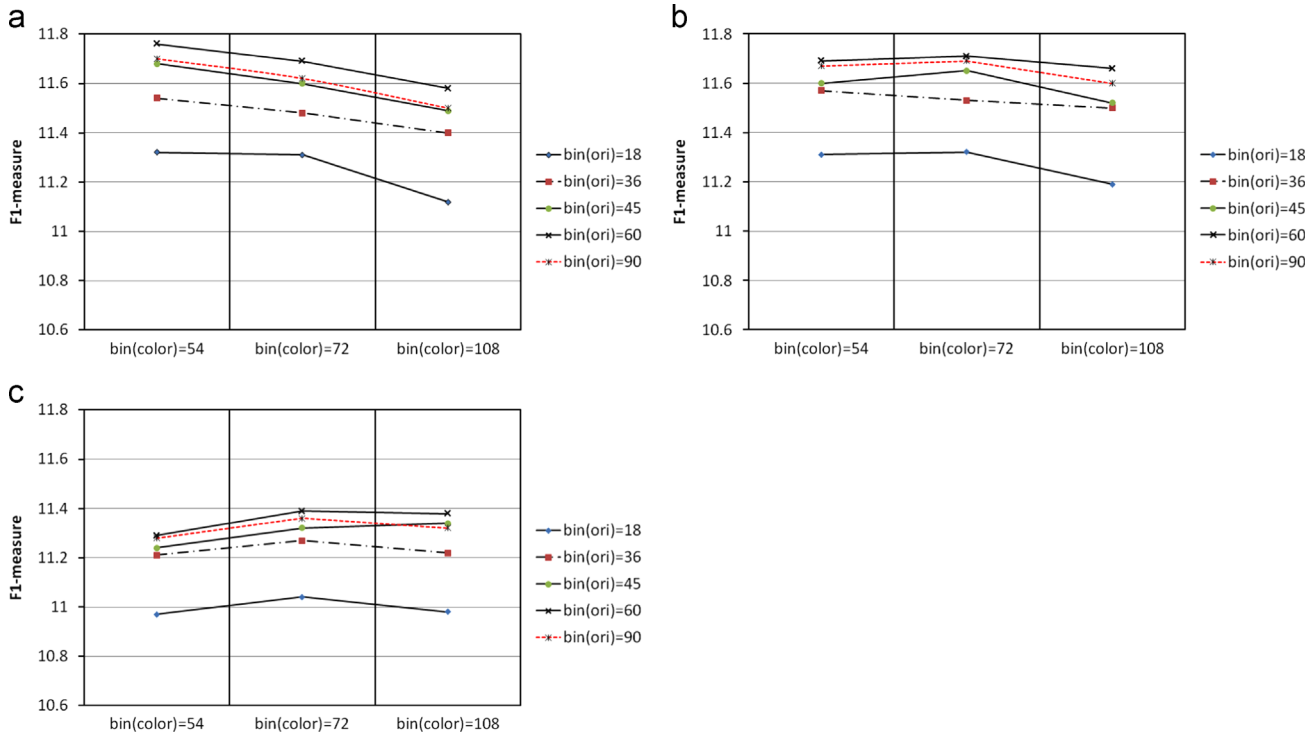

Fig. 6. The $F1$ measure of using saliency structure histogram on Corel-10K dataset, where the quantization levels of intensity are fixed as 16 bins, 32 bins and 64 bins, respectively: (a) the quantization level of intensity is 16, (b) the quantization level of intensity is 32, and (c) the quantization level of intensity is 64. Where the quantization level of orientation is denoted as bin(ori), the quantization level of color is denoted as bin (color).

Fig. 7. A retrieval example of using saliency structure histogram on Corel-10K dataset. The query is a polo image, and 11 returned 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.)

In order to give the best trade-off between the performances and vector dimensionality, the final quantization levels of color, edge orientation, and intensity are set as 54, 60 and 16 bins, respectively. Thus, the final vector dimension is $54+60+16=130$ in image retrieval. In this case, the best performance of the proposed algorithm is achieved.

Fig. 7 shows a retrieval example on Corel-10K dataset. The example only used to validate what visual characteristics may be reflected in the proposed algorithm, and does not to suggest that all queries on Corel-10K dataset can obtain such high retrieval accuracy. In Fig. 7, the query is a polo image, and 11 retrieved images show good match of shape, texture and color to the query image. Accordingly, the proposed algorithm has the discrimination power of color, texture and edge features. In Fig. 7, there only an image is not belong to the category of polo, but all the top 12 retrieved images have horse and people.

5.4.2. Varying the mask size of Gabor filters

As mentioned before, Gabor energy of the saliency structures also control the performance of the proposed algorithm. The goal here is to evaluate the performance of the proposed algorithm with different mask sizes in the computation of oriented Gabor filters. The computation of oriented Gabor filters is already described in Section 4.3, and the results are summarized in Table 1.

As can be seen from Table 1, the performance of saliency structure histogram is slowly reducing with mask size increase, and the computational burdens also increase. In other words, mask size has slight influence on the final retrieval performance, meanwhile the computational burdens increase. Thus, 9×9 mask is more suitable for the computation of oriented Gabor filters in the proposed algorithm.

5.4.3. Evaluation of different primary visual features

It is well known that different primary visual features provide different contribution to retrieval performance. A certain feature may be strong in one case but weak in another. In order to investigate the contribution of combining different primary visual features, we have implemented different combinations of color, intensity and edge orientation on Corel-10K dataset, where the quantization level of color, orientation and intensity are 54, 60 and 16, respectively.

Table 1
The performance of saliency structure histogram with different mask sizes in the computation of oriented Gabor filters on Corel-10K dataset.

Datasets	Performances	Different mask sizes				
		9×9	11×11	13×13	15×15	17×17
Corel-10K	Precision (%)	54.88	54.80	54.75	54.53	54.53
	Recall (%)	6.58	6.57	6.57	6.54	6.54
	*F*1-measure (%)	11.76	11.74	11.73	11.69	11.69

As can be seen from Fig. 8, two-fold can be mainly summarized: (1) in many cases, only using single visual feature cannot obtain good performance. In the proposed algorithm, the performance of color information outperforms that of orientation and intensity, where only using orientation achieves the worst performance. (2) Combining two or three visual features can significantly improve retrieval performances, where the combinations of color and intensity or orientation achieve better performances than the combination of orientation and intensity.

In summary, the performances are significantly increased by combining three feature types, instead of using only one of them or two of them.

5.4.4. Evaluation of color-volume and bar-shaped structures

As mentioned before, color-volume and bar-shaped structures are introduced into the proposed saliency model within CBIR framework. In practice, two cues provide different contribution to improve retrieval performance. In order to investigate the contribution of color-volume and bar-shaped structures, we have implemented different combinations of color-volume and bar-shaped structures on Corel-10K dataset and GHIM-10k dataset.

As can be seen from Table 2, without using color-volume or bar-shaped structures, the performance reduce when compared to

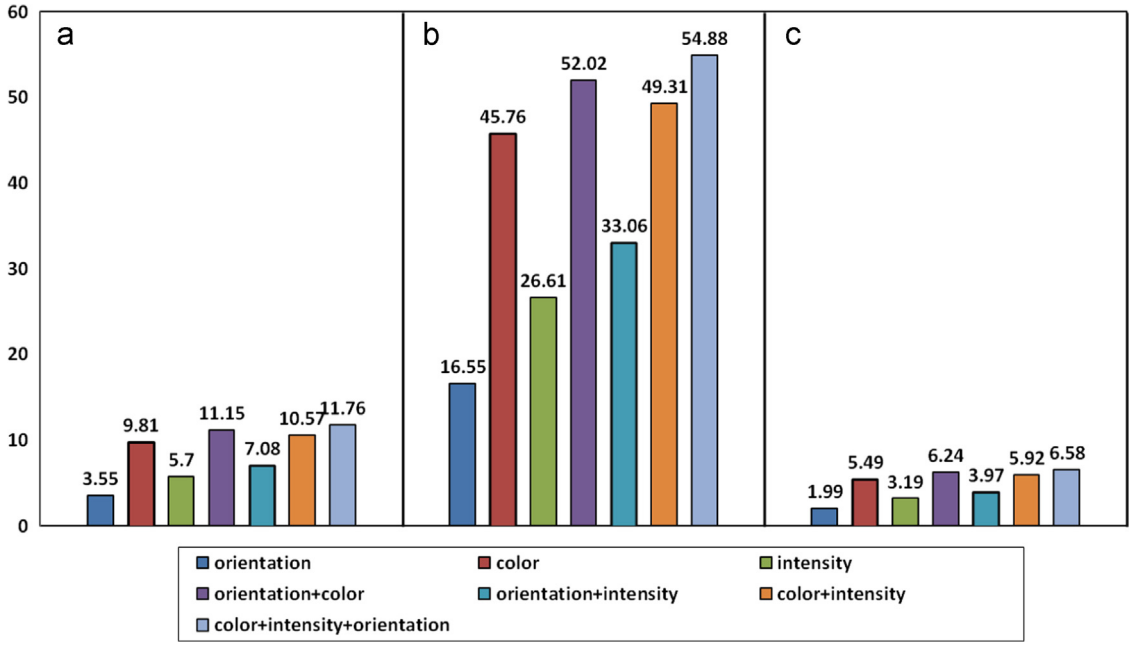

Fig. 8. The retrieval performance of combining different primary visual features on Corel-10K dataset: (a) *F*1-measure(%), (b) precision(%) and (c) recall (%).

Table 2
The performance of saliency structure histogram using color-volume or bar-shaped structures.

Dataset	The setting of color-volume or bar-shaped structures	Performances		
		Precision (%)	Recall (%)	*F*1-measure (%)
Corel-10K	Without color-volume	53.10	6.37	11.38
	Without bar-shaped structure	46.92	5.63	10.06
	Default setting	54.88	6.58	11.76
GHIM-10K	Without color-volume	60.56	1.45	2.83
	Without bar-shaped structure	55.21	1.33	2.59
	Default setting	61.16	1.47	2.87

default setting. Without using bar-shaped structures, the performances significantly reduce on two datasets. It is clear that, using both of color-volume and bar-shaped structures can improve the performances of saliency structures histogram, where the bar-shaped structures have the biggest contribution in the improvement of performances. As can be seen from Fig. 3(b) and (c), the color conspicuity areas can be straightforward pop-out by using the color volume information, and suppressed the background to certain extent. Such an attribute of color-volume is helpful to detect saliency areas, thus it still has a lot of room to study on how to use color-volume in future works.

5.4.5. Performance comparisons among BOW, MSD and the proposed algorithm

Fig. 9 summarizes the performances among the standard BOW baseline, micro-structure descriptor and saliency structure histogram. On Corel-10K and GHIM-10K datasets, the proposed algorithm achieves better performances than the standard BOW baseline and MSD. The average values of precision, recall and *F*1-measure on the two datasets are listed in Fig. 9. It should be stressed that the proposed algorithm can obtain such performances only with $54+60+16=130$ dimension vector, while the vector dimension of the standard BOW baseline is 1000, and its vector dimension beyond that of MSD and SSH algorithm very much.

As can be seen from Fig. 9, there are significant differences in the retrieval performances among the standard BOW baseline, MSD and SSH. On Corel-10K dataset and GHIM-10K dataset, the images of each category have similar contents, and only a small part of the images have contained the same object or scene viewed under different imaging conditions, thus, the standard BOW baseline does not achieve good performances. It is clear that BOW algorithm is unsuitable for content-based image retrieval. However, it does not mean that the proposed algorithm will be better than BOW algorithms in objects retrieval or objects categorization. It only validates that the proposed algorithm is suitable for content-based image retrieval, and BOW algorithms are suitable for object-based image retrieval or objects recognition.

5.4.6. Discussions

It is very worth noting that the differences among object recognition, object-based image retrieval and content-based image retrieval (CBIR), because this paper is limited to the scope of content-based image retrieval. The differences can be mainly summarized as three-fold: (1) Object recognition focus on finding and identifying objects in an image or video sequence. It is very important that the training dataset and test dataset must be kept distinct. (2) Object-based retrieval focus on searching for the same object as the target object. It is one example of searching the images containing same object or scene viewed under different imaging conditions [34,38,39]. Generally speaking, object-based retrieval has no distinction between the training dataset and test dataset. (3) However, CBIR focus on searching for the similar images as the given one. In many cases, content-based image retrieval is so called similar image retrieval.

In summary, the most important distinction among the three techniques is to find whether the same object or similar images.

In our earlier work [3], micro-structures are defined as the collection of certain underlying colors which have similar or the same edge orientation in uniform color space. There are two major advantages of micro-structure model: (1) it can mimic human's color perception well, (2) it can serve as a bridge to combine the color, texture and shape features as a whole. However, orientation-selective mechanism and saliency information are discarded. In

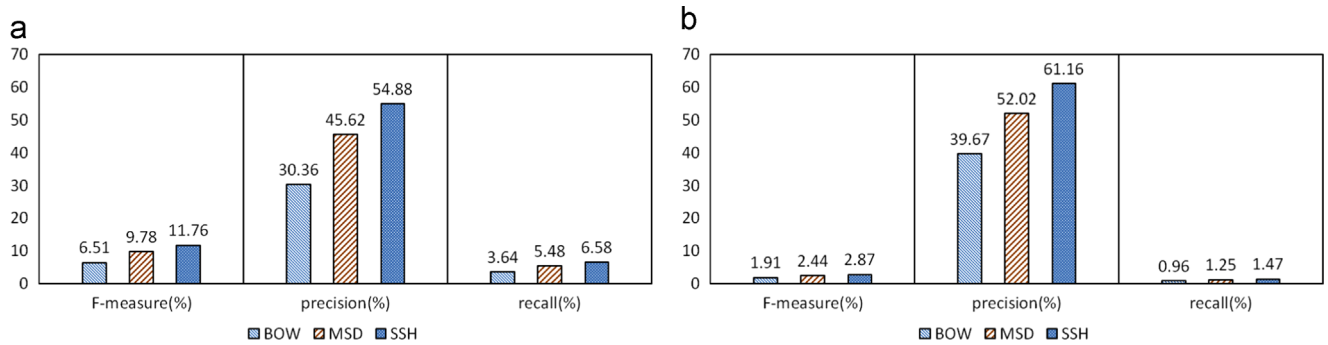

Fig. 9. Performance comparison among the standard BOW baseline, micro-structure descriptor (MSD) and saliency structure histogram (SSH) on two datasets: (a) Corel-10K dataset and (b) GHIM-10K dataset.

Fig. 10. A retrieval example using saliency structure histogram on GHIM-10K dataset. The query is a motorcycle image, and all returned 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.)

summary, micro-structures model has not simulated the visual attention mechanism well.

In the proposed model, the combination of color volume and edge information is used to detect saliency regions instead of using the primary visual features (e.g., color, intensity and orientation). Thus, part of background is shielded and can pop out the major objects in the images. It can provide a helpful complement to describe image contents, leading to the improvement of retrieval performance. Moreover, bar-shaped structures not only contain the spatial arrangements of the same-valued pixels but also have a sense of direction. The spatial arrangements and direction-sense are very important factors in visual attention. Combining bar-shaped structures with oriented Gabor filters can simulate orientation-selective mechanism well.

In image representation, the scheme of saliency structure histogram is to generate three tuples histograms considering the bar-shaped structures and oriented Gabor filters via a very special type, where color, edge orientation and intensity information are mapped into histogram. In this manner, color, texture and edge features have combined into a cohesive whole. Fig. 10 shows a retrieval example using saliency structure histogram on GHIM-10K dataset. In Fig. 10, all retrieved images have similar color, texture and shape features. It should be emphasized that orientation-selective mechanism has reflected in our model. It is also the most significant progress of saliency structure histogram compared to micro-structure descriptor, where oriented Gabor filters are embedded into bar-shaped structures, and the Logarithm characteristic of Gabor energy is used as the values of histogram, thus, the discrimination power of saliency structure histogram is stronger than that of micro-structure descriptor, leading to the better retrieval performance [3].

In the standard BOW baseline, local features extracted from an image by using SIFT, SURF or other keypoints detectors, and then mapped into a set of visual words. Finally, an image is represented as a histogram of visual word occurrences. The local descriptors (e.g., Sift, SURF, ORB and other local descriptors) are originally developed for performing reliable matching between different views of an object or scene. Thus, BOW models are widely used in objects recognition, objects-based image retrieval and scene categorization, and can achieve very good performance [34–41]. Unfortunately, visual words obtained by using the vector quantization of local features descriptors which need heavy computational burdens and may result in the loss

information. Besides, spatial information or semantic attributes are not embedded into the standard BOW baseline, and therefore satisfactory results are not obtained on Corel-10K dataset and GHIM-10K dataset. Accordingly, it is suitable for object-based image retrieval rather than CBIR.

In BOW models, the number of clusters K is considered as an important parameter, which can significantly affect the performances of objects retrieval. A certain vocabulary size may be strong in one case but weak in another. In this paper, $K=1000$ is adopted in the standard BOW baseline. It is clear that the dimensional vector of the standard BOW baseline is very high.

Based on the above analysis, we have known that the performances of the proposed algorithm are beyond the standard BOW baseline and MSD in the CBIR experiments, the standard BOW baseline is unsuitable for content-based image retrieval. Even so, BOW model is considered as one of state-of-the-art methods in objects recognition and scene categorization, and can obtain excellent performances on PASCAL VOC 2007 dataset, Caltech-101 dataset and other datasets using for identifying an object. Besides, the standard BOW baseline is also considered as an efficient visual search method of videos cast [39]. Thus, BOW techniques can be extended in future work to our saliency model within CBIR framework.

6. Conclusion

In this paper, we have developed a novel computational visual attention model to improve the performances of our earlier work (namely micro-structure descriptor) for content-based image retrieval. First, a novel visual cue, namely color volume, with edge information together is introduced to detect saliency regions instead of using the primary visual features (e.g., color, intensity and orientation). Second, the energy feature of the gray-level co-occurrence matrices is used for globally suppressing maps, instead of the local maxima normalization operator in Itti's model. At last, oriented Gabor filters are embedded into bar-shaped structure to simulate orientation-selective mechanism and image representation.

The saliency structures histogram considers the bar-shaped structures and oriented Gabor filters via a very special type, where color, edge orientation and intensity information are mapped into histograms. In the proposed algorithm, oriented Gabor filters can be embedded into bar-shaped structures, and the Logarithm characteristic of Gabor energy is used as the values of histograms. Thus, it has good discrimination power of color, texture, edge feature and spatial layout. The proposed algorithm can be considered as an improved version of our earlier work (micro-structure descriptor) by combining a bottom-up component of visual attention and orientation-selective mechanism.

The experimental results on Corel-10K dataset and GHIM-10K dataset have demonstrated that it is much more efficient than representative image feature descriptors, such as the standard BOW baseline and micro-structure descriptor in content-based image retrieval.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61233011, 61202272, 61463008, 61202318, and 61363035). The authors would like to thank the anonymous reviewers for their constructive comments.

References

[1] D. Hubel, T.N. Wiesel, Receptive fields. Binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1962) 106–154.
[2] E.H. Adelson, J.R. Bergen., The plenoptic function and the elements of early vision, in: M. Landy, J. Movshon (Eds.), Computational Models of Visual Processing, MIT Press, Cambridge, 1991, pp. 3–20.
[3] G.-H. liu, Z.-Y. Li, L. zhang, Y. Xu, Image retrieval based on micro-structure descriptor, Pattern Recognit. 44 (9) (2011) 2123–2133.
[4] A. Treisman., A feature in integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[5] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[6] J.M. Wolfe, T.S. Horowitz, What attributes guide the deployment of visual attention and how do they do it? Nat. Rev. Neurosci. 5 (6) (2004) 495–501.
[7] J.K. Tsotsos, S.M. Cuihane, et al., Modeling visual attention via selective tuning, Artif. Intell. 78 (1) (1995) 507–545.
[8] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Netw. 19 (9) (2006) 1395–1407.
[9] O.L. Meur, P.L. Callet, et al., A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.
[10] Y. Sun, R. Fisher, Object-based visual attention for computer vision, Artif. Intell. 20 (11) (2003) 77–123.
[11] A. Borji, L. Itti. Exploiting local and global patch rarities for saliency detection, in: 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 478–485.
[12] A. Toet, Computational versus psychophysical bottom-up image saliency: a comparative evaluation study, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2131–2146.
[13] A. Borji, L. Itti., State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 185–207.
[14] T. Kadir, M. Brady, Saliency, scale and image descriptor, Int. J. Comput. Vis. 45 (2) (2001) 83–105.
[15] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, John Wiley & Sons Ltd., New York, 2002.
[16] R.M. Haralick, Dinstein Shangmugam, Textural feature for image classification, IEEE Trans. Syst. Man Cybern. SMC-3 (6) (1973) 610–621.
[17] H. Tamura, S. Mori, T. Yamawaki, Texture features corresponding to visual perception, IEEE Trans. Syst. Man Cybern. 8 (6) (1978) 460–473.
[18] G. Cross, A. Jain, Markov random field texture models, IEEE Trans. Pattern Anal. Mach. Intell. 5 (1) (1983) 25–39.
[19] B.S. Manjunathi, W.Y. Ma., Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 837–842.
[20] T. Ojala, M. Pietikanen, T. Maenpaa, Multi-resolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[21] G.-H. Liu, J.-Y. Yang, Image retrieval based on the texton co-occurrence matrix, Pattern Recognit. 41 (12) (2008) 3521–3527.
[22] G.-H. Liu, L. Zhang, et al., Image retrieval based on multi-texton histogram, Pattern Recognit. 43 (7) (2010) 2380–2389.
[23] J. Luo, D. Crandall, Color object detection using spatial-color joint probability functions, IEEE Trans. Image Process. 15 (6) (2006) 1443–1453.
[24] G.-H. Liu, J.-Y. Yang, Content-based image retrieval using color deference histogram, Pattern Recognit. 46 (1) (2013) 188–198.
[25] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory (1962) 179–187.
[26] P.-T. Yap, R. Paramesran, S.-H. Ong., Image analysis using Hahn moments, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 2057–2062.
[27] F.P. Kuhl, C.R. Giardina, Elliptic Fourier features of a closed contour, Comput. Graph. Image Process 18 (1982) 236–258.
[28] C.T. Zahn, R.Z. Roskies, Fourier descriptors for plane closed curves, IEEE Trans. Comput. C-21 (3) (1972) 269–281.
[29] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[30] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2, 2004, pp. 506–513.
[31] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[32] H. Bay, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, in: Proceedings of the European Conference on Computer Vision 1, 2006, pp. 404–417.
[33] K. Mikolajczyk, T. Tuytelaars, C. Schmid, et al., A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1–2) (2005) 43–72.
[34] J..Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: Proceedings of IEEE International Conference on In Computer Vision, 2003.
[35] D. Nister, H. Stewenius. Scalable recognition with a vocabulary tree, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2006, pp. 2161–2168.
[36] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[37] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of features image classification, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 490–503.
[38] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: improving particular object retrieval in large scale image databases, in:

Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[39] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 591–606.
[40] Y. Su, F. Jure, Improving image classification using semantic attributes, Int. J. Comput. Vis. 100 (1) (2012) 59–77.
[41] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.
[42] P.L. Rosin, Measuring corner properties, Comput. Vis. Image Underst. 73 (2) (1999) 291–307.
[43] M. Partio, B. Cramariuc, M. Gabbouj, A. Visa. Rock texture retrieval using Gray level co-occurrence matrix, in: Proceedings of the 5th Nordic Signal Processing Symposium, Norway, 2002.
[44] W. Burger, M.J. Burge, Principles of Digital image processing: Core Algorithms, Springer, New York, 2009.
[45] R.C. Gonzalez, R.E. Woods, Digital Image Processing, 3rd edition, Prentice-Hall, New York, 2007.
[46] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, Opt. Soc. Am. 4 (12) (1987) 2379–2394.
[47] J. Mutch, D.G. Lowe, Object class recognition and localization using sparse features with limited receptive fields, Int. J. Comput. Vis. 80 (1) (2008) 45–57.
[48] S.C. Dakin, R.J. Watt., The computation of orientation statistics from visual texture, Vis. Res. 37 (22) (1997) 3181–3192.
[49] B. Julesz., Textons, the elements of texture perception and their interactions, Nature 290 (5802) (1981) 91–97.
[50] B. Julesz, Texton gradients: the texton theory revisited, Biol. Cybern. 54 (1986) 245–251.
[51] A.J. Bell, T.J. Sejnowski., The 'independent components' of natural scenes are edge filters, Vis. Res. 37 (23) (1997) 3327–3338.
[52] J. Malik, P. Perona, Preattentive texture discrimination with early vision mechanisms, J. Opt. Soc. Am. A 7 (1990) 923–932.
[53] D. Marr, E. Hildreth, Theory of edge detection, Proc. R. Soc. Lond. Ser. B: Biol. Sci. 207 (1167) (1980) 187–217.
[54] L.W. Renninger, J. Malik, When is scene identification just texture recognition, Vis. Res. 44 (19) (2004) 2301–2311.
[55] C.J. van Rijsbergen, Information Retrieval (2nd end), Butterworths, London, 1979.
[56] G. Hripcsak, A.S. Rothschild, Agreement, the *f*-measure, and reliability in information retrieval, J. Am. Med. Inf. Assoc. 12 (3) (2005) 296–298.
[57] S.H. Schwartz, Visual Perception: A Clinical Orientation, fourth edition, McGraw-Hill, New York, NY, 2009.
[58] C.F. Hall, E.L. Hall, A nonlinear model for the spatial characteristics of the human visual system, IEEE Trans. Syst. Man Cybern. 7 (3) (1977) 161–170.
[59] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: Proceedings of IEEE International Conference on In Computer Vision, 2011, pp. 2564–2571.
[60] C. Harris, M. Stephens. A combined corner and edge detector, in: Alvey Vision Conference, 1998, pp. 147–151.

Guang-Hai Liu is currently an 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). In 2011, He was engaged as an evaluation expert of science and technology project of Guangxi, China. His current research interests are in the areas of image processing, pattern recognition and artificial intelligence.

Jing-Yu Yang received the B.S. Degree in Computer Science from the 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.

Zuo-Yong Li received the B.S. degree in computer science and technology from the Fuzhou University in 2002. He got his M.S. degree in computer science and technology from the Fuzhou University in 2006. Now, he is a Ph.D. Candidate in the Nanjing University of Science and Technology and an associate professor in the Department of Computer Science of Minjiang University. He has published several papers in international/national journals. His research interests include image segmentation and pattern recognition.