Multimed Tools Appl (2018) 77:29099–29117
https://doi.org/10.1007/s11042-018-6122-2

Content-based image retrieval via a hierarchical-local-feature extraction scheme

Muwei Jian[1] · Yilong Yin[2] · Junyu Dong[3] · Kin-Man Lam[4]

Received: 21 September 2017 / Revised: 14 March 2018 / Accepted: 8 May 2018 /
Published online: 16 May 2018

Abstract Recently, with the development of various camera sensors and internet network, the volume of digital images is becoming big. Content-based image retrieval (CBIR), especially in network big data analysis, has attracted wide attention. CBIR systems normally search the most similar images to the given query example among a wide range of candidate images. However, human psychology suggests that users concern more about regions of their interest and merely want to retrieve images containing relevant regions, while ignoring irrelevant image areas (such as the texture regions or background). Previous CBIR system on user-interested image retrieval generally requires complicated segmentation of the region from the background. In this paper, we propose a novel hierarchical-local-feature extraction scheme for CBIR, whereas complex image segmentation is avoided. In our CBIR system, a perception-based directional patch extraction method and an improved salient patch detection algorithm are proposed for local features extraction. Then, color moments and Gabor texture features are employed to index the salient regions. Extensive experiments have been performed to evaluate the performance of the proposed scheme, and experimental results show that the developed CBIR system produces plausible retrieval results.

Keywords Network big data · Content-based image retrieval · Perception-based directional patch · Salient patch detection

✉ Muwei Jian
jianmuweihk@163.com

[1] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China

[2] School of Software Engineering, Shandong University, Jinan 250101, China

[3] Department of Computer Science and Technology, Ocean University of China, Qingdao, China

[4] Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

1 Introduction

In the big data era, content-based image retrieval (CBIR) has been a hot research topic in the last few years, because of its many practical applications, including in the fields of multimedia management and analysis [5, 40], biomedical image retrieval [43], dermatological lesions analysis [21], intelligent video surveillance [15, 18, 19, 49], traffic congestion analysis [48, 51], videos captioning [10], and so on.

Nowadays, with the progress of various camera sensors, multimedia technology and internet network, the number of digital images is increasing rapidly [55]. These vast and tremendous amounts of big-image-data have important applications in both military and civil domains. Therefore, how to effectively and efficiently manage and retrieve image from internet network has became a hot research issue. Traditional image search/query technology by text (QBT) mainly depends on manual labeled image. However, QBT has obvious disadvantages and shortcomings: different people has various understanding on the identical image content, and manual-based labeling has result in that it is unpractical to label with the great variety of images in the internet network. As a result, those search engines produced match results are not actually equivalent to what users are really seeking or interested. In 1990's, content-based image retrieval (CBIR) was firstly proposed and investigated. Content-based image retrieval uses the image content itself, such as color, shape and texture features, instead of using the manual labels of images as the search index [17, 25]. The CBIR system extracts the query image's features (Query by Example), and retrieves from the internet network or the image database.

In general, one of the main steps in CBIR system is how to effectively extract feature vectors that can be utilized to depict the relevant visual contents of images [27]. Then, feature vectors of the given input example (Query image) and the candidate images are compared based on a mathematical similarity function, and the images containing approximately close feature vectors are returned. Nevertheless, the traditionally CBIR methods using holistic/global feature vectors normally cannot represent all the characteristics of the image, the reason is that a typical image also involves local features that are associated with some objects or certain meaningful regions of interest. Distinctly, local features are crucial and useful to depict an image in CBIR system [17, 26, 53]. To bridge the semantic gap, a generic semantic image/video indexing method based on three levels of analyzing the spatial information in visual content was proposed in [6]. Recently, Xu designed a multiple-instance learning (MIL) based decision neural network that attempts to learn the concept from the training bags for CBIR [52]. Interested readers can refer to the survey paper on semantic content-based image retrieval [1].

In addition, Human psychology discovers that people commonly pay close attention to the visual meaningful regions of their interest in an image [20, 30]. Thus, in a CBIR system, users only want to search images containing similar regions to the query example, whereas neglecting unrelated image areas (such as the irrelevant texture regions or the image background). A straightforward method to solve this issue is to utilize image segmentation algorithms to segment the consumer's concerned regions or objects from the background. Previous relevant literatures [2, 8, 9, 23, 44] are published on this issue and a large number of segmentation-based algorithms are introduced to discuss the local features representation for CBIR. The obvious drawback of the segmentation-based CBIR systems is that, when the result of image-segmentation procedure is not precise enough, the experimental results will be seriously deteriorated. What's more, when segmentation failure occurs during the search process, it will be produce even inferior retrieval performance [25]. To sum up, the precise and general image-segmentation technique is still an open and hard issue in the domain of computer vision and pattern recognition. Consequently, as suggested in [31] by Pavlidis, it would be difficult to pursue general CBIR when we do not have robust segmentation methods that can identify objects on an image efficiently.

In order to exploit saliency region for CBIR, different saliency detection models have been investigated. In [46], Wang et al. proposed a novel supervised Markov-random-field framework based on differential threshold for saliency detection. Based on multiple-instance learning, a saliency detection model by integrating low-, mid-, and high-level features in the detection procedure was proposed [47]. Meanwhile, for scalable image and video retrieval, a general quantization-based hashing framework, which preserves the properties in the original space and reduces the quantization error, was proposed in [38]. To tackle the bottleneck of matching speed in large-scale dataset, Song et al. [39] designed an efficient Deep Region Hashing (DRH) method using an image patch as the query. Interested investigators can refer to the comprehensive survey paper on learning to hash [50].

Since psychological investigation observed that a local region with high global variations normally cause human's concern and visual attention, which plays a vital role in the designing of effective CBIR systems. In this paper, we develop a novel hierarchical-local-feature extraction approach in a biologically and psychological plausible way for CBIR, which can keep away from complicated and time-consuming image segmentation. The main advantages and contributions of this paper are as follows:

(1). Regarding that directionality information is one of the important and intrinsic features for the human visual system (HVS) to perceive and analyze images, a perception-based directional patch extraction method is proposed to describe the local features of an image for CBIR, while the global image features can be avoid. By using wavelet decomposition, the directionality patches in an image, from the corresponding vertical, horizontal, and diagonal wavelet sub-bands, can be detected and extracted.

(2). Simultaneously, the visually salient patches can always catch human's visual attention and interest [30]. Therefore, an improved wavelet-transform salient patch detector based on Chauvenet's criterion is designed. These extracted patches can be utilized as the local properties of the image to characterize an image, so that the global image features and image segmentation can also be excluded.

(3). Both the proposed two local feature extraction methods, aiming to extract and capture the image's local characteristics, have an implicit merit, i.e., since psychological research has shown that the HVS applies a multi-scale frequency analysis when we analyze an image [20], the wavelet-decomposition process is closely resembled to the zooming-in and the zooming-out mechanism of the HVS.

(4). A clustering algorithm is performed to parallelly divide the extracted directional and salient patches into different regions in terms of their spatial distribution. The clustering scheme in the proposed CBIR system can divide the salient patches into meaningful subclasses, which can be further used as a hierarchical-local-feature during the image-retrieval process.

(5). A hierarchical-local-feature extraction scheme comprising three levels is devised to represent an image, which is particularly suitable for network big-image-data index and management. In the proposed approach, both the global and local features are taken into consideration simultaneously, whereas time-consuming and complex image segmentation is avoided.

The rest of the paper is organized as follows. In section 2, we will introduce the proposed hierarchical-local-feature extraction scheme, and present the clustering algorithm according to the spatial distribution of the extracted patches in detail. Experimental results and comparisons are given in section 3. Finally, the paper draws a conclusion in section 4.

2 Hierarchical-local-feature extraction scheme

The framework of the proposed novel scheme based on a three-level hierarchical structure for content-based image retrieval is illustrated in Fig. 1.

In the first level, global features, i.e. color moments [11] and Gabor filters features [13, 36, 54], of the whole image are firstly utilized to form a global/holistic feature vector to

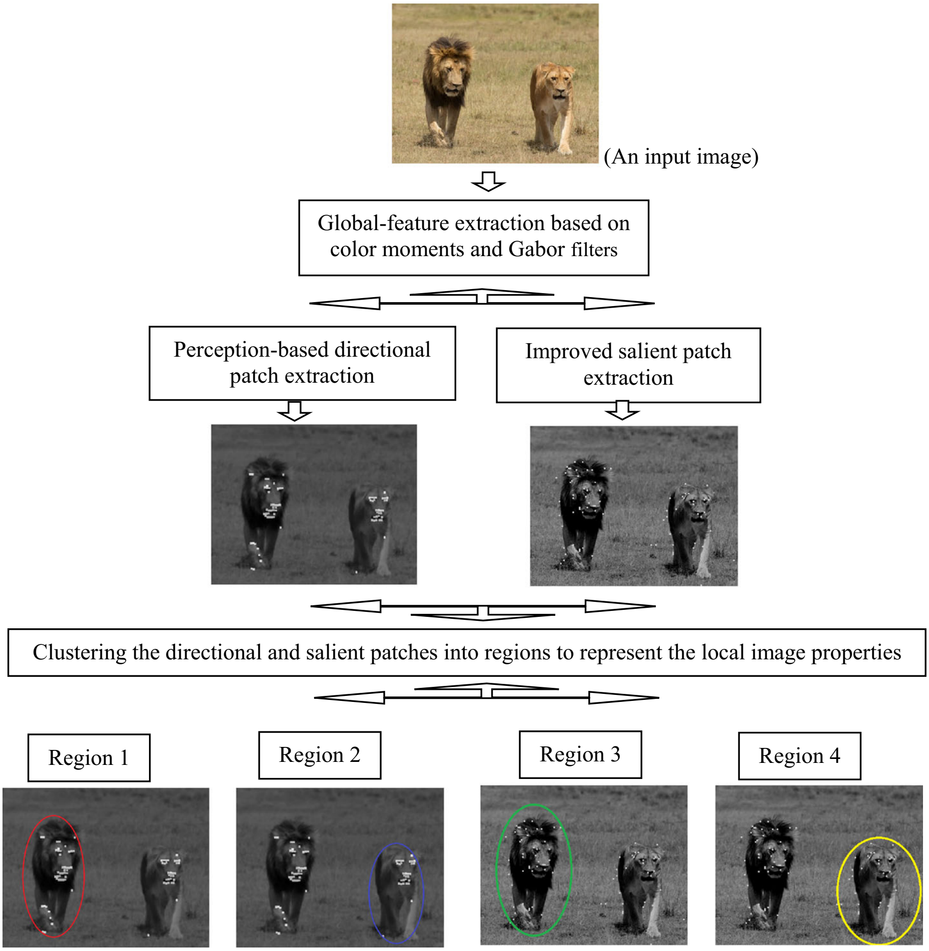

(An input image)
Global-feature extraction based on color moments and Gabor filters
Perception-based directional patch extraction
Improved salient patch extraction
Clustering the directional and salient patches into regions to represent the local image properties
Region 1
Region 2
Region 3
Region 4

Fig. 1 The proposed framework for CBIR based on a three-level hierarchical structure. The different color circles represent the different meaningful regions to represent the local image properties, which are clustered into different meaningful subclasses based on the directional- and the visually salient patches of the input image, accordingly. Finally, these different meaningful regions (leaf nodes of the TREE Hierarchical data structure) are also utilized to construct local feature vectors for describing the separated part of the image to carry out CBIR

index the image. Next, in the second level, directional- and salient-patches detectors are developed to extract the perception-based directional patches and the visually salient patches of the image in parallel. In the third phase, a clustering algorithm is performed to divide the two typical patches (e.g. directional and salient patches) into different and meaningful regions/subclasses to characterize the local image characteristics. Finally, the different meaningful regions, namely leaf nodes of the TREE Hierarchical data structure, are also utilized to construct local feature vectors for describing the separated part of the image to carry out CBIR.

The following sections will introduce the details of the proposed hierarchical-local-feature extraction structure.

2.1 Perception-based directional patch extraction

In human history, sense of direction-perception has been born in the course of society development, and has developed its concept of directionality/orientation. Therefore, directionality/orientation cue in an image is a solid and credible feature, and can be used this intrinsic feature for the human visual system (HVS) to perceive and analyze image. Thereby, the directionality/orientation feature is very beneficial and helpful for some computer-vision applications, for example saliency detection [20] and image retrieval [16].

In this subsection, we will use a wavelet-based algorithm to extract directional patches. The wavelet transform, which belongs to the multi-resolution analysis (MRA) technique, can present and reflect image variations at different resolutions and scales [4, 28]. By the definition of wavelet theory, a wavelet is an attenuated and oscillating function, and defining the two-dimensional discrete wavelet transform (DWT) of the image $f(x, y)$ can be written [4, 28]:

$$\begin{gathered}\mathrm{W}_{\varphi}(\mathrm{j}_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f \varphi_{\mathrm{j}_0, m, n} \text{ and} \\ \mathrm{W}_{\psi}{}^{\mathrm{i}}(\mathrm{j}_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f \psi^{i}_{\mathrm{j}_0, m, n},\end{gathered} \tag{1}$$

where $M \times N$ denotes the size of $f(x, y)$; j_0 is an arbitrary initial scale; $\varphi(x)$ is the scaling function, and $\psi(x)$ is the wavelet function; and $i = \{H, V, D\}$, where H is the horizontal high-frequency information, V the vertical high-frequency information, and D the diagonal high-frequency information, respectively.

Since the $\{H, V, D\}$ sub-images of wavelet transform cover three different direction high-frequency information, at first, we calculate the vertical, horizontal, and diagonal directionality maps of the three $\{H, V, D\}$ sub-images. Then, these $\{H, V, D\}$ directionality maps are re-scaled and normalized to integrated into a compound directionality map for detecting the perceptual-meaningful directional patches.

With regard to horizontal directionality computation, which is computed from the HL sub-band (including the vertical high-frequency information) of the wavelet transform. Suppose $h(x, y)$ is the sub-band coefficients of the wavelet transform, where x and y denote the row coordinate and column coordinate of the sub-band, accordingly. Regarding every row of the HL sub-band image, the horizontal directionality is expressed in the following form [20]:

$$Dir_H(x, y) = \frac{1}{2W} |h(x, y)| \left\{ \sum_{r=-W}^{W} |h(x + r, y)| \right\}, \tag{2}$$

where W represents the width of adjacent coefficients to be extracted, and in practice it can be set to p (where p is the wavelet regularity of a orthogonal wavelet with a compact support).

In allusion to the computation of horizontal directionality, the vertical directionality is defined along each column:

$$Dir_V(x,y) = \frac{1}{2W}|h(x,y)|\left\{\sum_{r=-W}^{W}|h(x,y+r)|\right\}, \tag{3}$$

Contrasted with the horizontal and vertical directionality, the calculation of diagonal directionality is more complicated. The two perpendicularly diagonal directions, namely $\pi/4$ and $3\pi/4$, are taken into consideration together. The calculation process bears a resemblance to those computation for the horizontal and vertical directionality. Then, the diagonal directionality is defined as the average of the two perpendicularly diagonal directions of $\pi/4$ and $3\pi/4$ in the form follows:

$$Dir_D(x,y) = \frac{1}{2}\left(Dir_{D1}(x,y) + Dir_{D2}(x,y)\right), \tag{4}$$

where Dir_{D1} and Dir_{D2} denotes the diagonal directionalities of $\pi/4$ and $3\pi/4$, respectively.

After fusing the vertical, horizontal and diagonal directionality maps, these three different directionality maps can be re-scaled to the same size of the original image, and then normalize and merge them to construct a compounded orientation/directionality map [20]. Then, these patches with first K (e.g. $K = 60$) largest values can be extracted from the orientation map. Figure 2b shows some detected perception-based directional patches with the maximum values extracted from the compounded directionality map. As illustrated in Fig. 2b, the proposed method can detect the directionally salient patches in the image efficiently, which can be seen as a typical local feature to represent the local properties of the image for CBIR.

2.2 Improved salient-patch extraction

In term of the wavelet decomposition theory, at a coarse scale, a local absolute maximum value of the wavelet-transform coefficients matches along with a region with high holistic variations and responses the singularity point of a signal. Consequently, a salient point detector based on wavelet transform, which can express image variations at different resolutions, is devised to detect the visually salient points [22–25].

A orthogonal compactly supported wavelet filter produces a non-redundant decomposition of a signal. For the compactly supported wavelets, every single wavelet coefficient $W_{2^j}f(n)$ at the scale 2^j can be represented with $2^{-j}p$ signal points, and p represents the wavelet regularity. Accordingly, we can further observe the wavelet coefficients at the finer scale 2^{j+1}. For the point in the scale 2^{j+1}, there is a set of coefficients reflected with the same point at the scale 2^j.

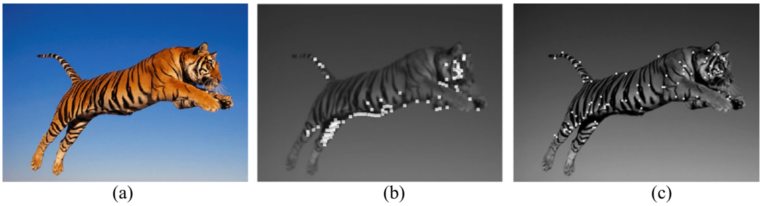

(a) (b) (c)

Fig. 2 Extraction of directional and salient patches: (**a**) an input image, (**b**) extraction of some directional patches, and (**c**) extraction of some salient patches, both are using the Db4 wavelet

These coefficients $C(W_{2^j}f(n))$ are called the children of the coefficient $W_{2^j}f(n)$, and the relationship between them is expressed in the following:

$$C(W_{2^j}f(n)) = \{W_{2^{j+1}}f(k), 2n \le k \le 2n + 2p - 1\}, \tag{5}$$

where $0 \le n \le 2^j N$ and N is the length of the signal.

In signal processing, the variation of a signal point can be reflected and calculated by their children coefficients $C(W_{2^j}f(n))$. Therefore, in an image, salient points can be detected and extracted by finding a relevant point to represent this global variation through examining coefficients at different scales. Classic methods employed the following formula which named saliency value to extract the salient points [33–35, 42]:

$$saliency = \sum_{k=1}^{-j} |C^{(k)}\left(W_{2^j}f(n)\right)|, 0 \le n \le 2^j N, -\log_2 N \le j \le -1. \tag{6}$$

However, the largest absolute values of the wavelet coefficients at different scales have different means and varying scope [14, 20].

Based on this observation, Jian and Dong proposed a revised salient point extraction algorithm in [14], which used a weight $w(k)$ (the reciprocal of the standard deviation) to normalized maximum wavelet coefficients to produce more accurate extraction results. Some details can be found in [14].

2.2.1 Improved salient-point extraction:

In this paper, for the purpose of detecting the salient points more precisely, an improved salient point detector is proposed firstly. According to statistical theory, the Chauvenet's criterion is a measurement of evaluating whether one piece of experimental data from a set of observations, is liable to be spurious [41]. Another method for eliminating spurious data is named "3σ" criterion which was introduced some years before Chauvenet's criterion was proposed. However, Chauvenet's criterion is a more rigorous method to the rational deletion of noise points and outlier data. In controversial practice, Chauvenet's criterion offers a quantitiative and objective approach for data rejection, particularly in small sets or where a Gaussian distribution cannot be assumed.

To employ Chauvenet's criterion, we should compute the mean $\overline{x}$ and standard deviation S_x of the experimental data firstly. Let n be the total number of the subband's wavelet coefficients, then Chauvenet's criterion can be simplified into:

$$\omega_n = 1 + 0.4\ln(n); \tag{7}$$

If $|x(j,k) - \overline{x}| < \omega_n * S_x$; let $x_{\max}$, $x_{\min}$ be the maximum and minimum value of subband's wavelet coefficients which satisfy$|x(j,k) - \overline{x}| > \omega_n * S_x$, respectively. We can use the following function (8) to normalize wavelet coefficients satisfying $|x(j,k) - \overline{x}| < \omega_n * S_x$ into [0, 1].

$$x'(j,k) = \frac{x(j,k) - x_{\min}}{x_{\max} - x_{\min}}. \tag{8}$$

Then, if $(x(j,k) - \overline{x}) \ge \omega_n * S_x$;
set $x'(j,k) = 1$;
If$(\overline{x} - x(j,k)) \ge \omega_n * S_x$;
set $x'(j,k) = 0$.

In order to weighted different scales and orientations subband of wavelet decomposition equally, the wavelet coefficient are processed by the Chauvenet's criterion. In this paper, the process of transforming the value into a range of 0 and 1 is called wavelet coefficients normalization.

Then the improved saliency values based on the normalized wavelet coefficients can be expressed:

$$Imp\_saliency = \sum_{k=1}^{-j} |C^{(k)}\Big(W'_{2^j} f(n)|, 0 \le n \le 2^j N, -\log_2 N \le j \le -1, \tag{9}$$

where wavelet coefficients in the three different orientations' and scales' subbands can be accumulated and weighted equally.

2.2.2 Extent improved salient-point to salient-patch extraction

Because the human visual ability and visual resolution is limited, in reality, people normally interest in these visually meaningful salient-patches of an image [14, 20] rather than the salient points. With the aim of reflecting the local characteristics of an image using those salient patches, a salient patch detector is developed to detect the visual meaningful patches in an image, and then represent this images' local characteristics for CBIR system. Therefore, we propose a method to extent the improved salient points detection to salient patches extraction for CBIR.

In detail, we define a set of pixels as the *s-patches* of the pixel *p*. By considering the wavelet coefficient at the scale with its *s-patches, w*e can further observe the *s-patches* at the finer scale. In other words, there is a set of coefficient calculated with the same points in the *s-patches* at the scale. These coefficients are named the children of the *s-patches*. Actually, the children coefficients reflect the variations of the *s-patches*, and the most salient patches are the sum of wavelet coefficients in *s-patches* with the largest absolute values at different scales. Accordingly, salient patches can be extracted by detecting this maximum and examining its largest child patches. The saliency at position (*x*, *y*) of an image is calculated by adding the saliency values inside Rect(*x*, *y*) in total. The whole saliency value of a patch is the summation of the normalized saliency values calculated from these three different sub-images. At last, we can detect the visually salient patches using the improved saliency values.

Figure 2c displays some visually salient-patches, which are detected using the improved salient-patch detection algorithm based on the Db4 wavelet. As illustrated in Fig. 2c, the developed salient-patch detector can extract the salient patches accurately, and can be used as a local feature to depict image's local characteristics for the proposed CBIR system.

2.3 Clustering using the spatial distribution of the extracted patches

In this subsection, we use a clustering algorithm to divide the two typical patches (e.g. directional and salient patches) into different and meaningful regions to represent the bottom-level-local image characteristics for CBIR.

As a popular and widely used method for clustering, the DBSCAN algorithm produces excellent performance in spatial data clustering. The DBSCAN clustering method was firstly investigated by Ester, et al. [7, 32], and the algorithm depends on a density-based notion of clusters. This clustering method is especially suited to deal with big data with noises. With only one input parameter, DBSCAN algorithm can cluster the data into different meaningful subclasses. Some details regarding the DBSCAN algorithm please refer to [7, 32].

In the proposed method, the DBSCAN algorithm is applied to identify clusters of salient patches and then apply the clusters to automatically divide salient patches into various meaningful salient regions. The basic idea of the DBSCAN algorithm is that, for every point of a cluster, the neighborhood of a given radius has to include at least a minimum number of points, namely, the density in the neighborhood has to exceed a predefined threshold. Thereby, the DBSCAN algorithm requires an input parameter - neighborhood radius, to divide salient patches into various meaningful subclasses in our CBIR system. In this paper, a simple algorithm is developed to determine the neighborhood radius threshold *Eps* adaptively:

Assume *K* salient patches that are extracted in an image, and for each salient patch, utilizing one center-point (x_c, y_c) to represent the spatial distribution of the salient patch accordingly. Let $Cen = (x', y')$ be the centroid of the center-point of the salient patches, and x' and y' are expressed as follows:

$$x^{'} = \frac{1}{K}\sum x_c; \tag{10}$$

$$y^{'} = \frac{1}{M}\sum y_c. \tag{11}$$

Define the centroid $Cen = (x', y')$ as the center of the detected salient patches. Let *R* be the maximum radius with the *Cen* as the center of a circle, it can be defined as follows:

$$R = \max\left[(x_c - x')^2 + (y_c - y')^2\right]^{\frac{1}{2}}. \tag{12}$$

After that, the neighborhood radius threshold can be determined:

$$\text{Mean\_}R = \pi R \times R/M; \tag{13}$$

$$Eps = \alpha^{*}\text{Mean\_}R; \tag{14}$$

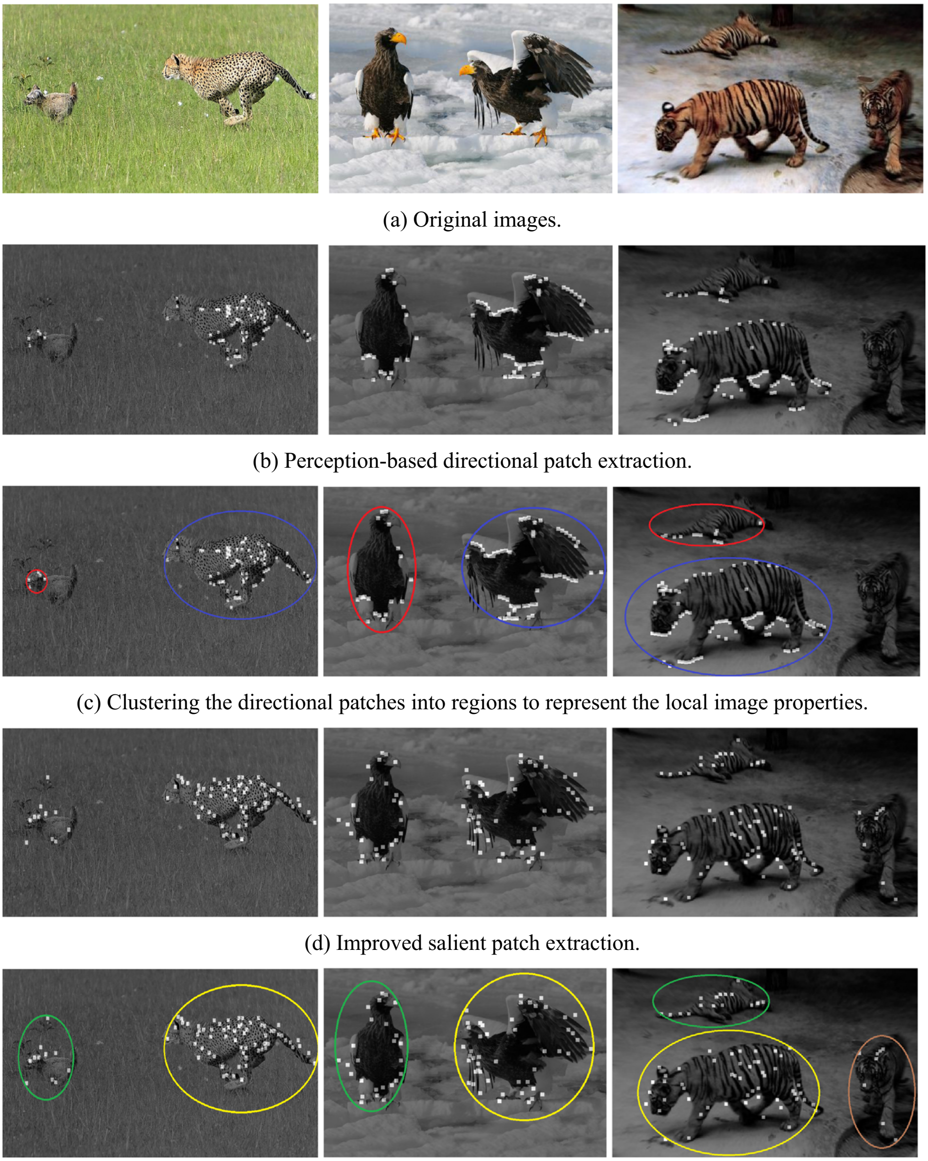
The parameter α can be set in the range of [0~1]. After the threshold *Eps* is obtained, we can localize the salient regions and efficiently cluster the salient patches into various meaningful subclasses, namely salient regions in terms of salient patches' spatial distribution to represent the bottom-level-local image characteristics for CBIR.

To sum up, the proposed scheme in our CBIR system is a three-level hierarchical structure, as shown in Fig. 1. In the first stage, color moments [11] and Gabor filters features [13] of the whole images in the database are utilized to construct a global/holistic feature vector to describe each image. In the second level, directional and salient patches detector are carried out to detect the perception-based directional patches and the visually salient patches of each image, accordingly. After the two typical patches (e.g. directional and salient patches) are extracted, color moments and Gabor texture features of the different typical patches are also utilized to build feature vectors for indexing the different patches of the image, called mid-level-local image feature, to reflect the local image properties (As shown in the second level in Fig. 1). In the third phase, the DBSCAN clustering algorithm is performed to cluster the two typical patches (e.g. directional and salient patches) into different and meaningful regions in terms of their spatial distribution, to represent the bottom-level-local image characteristics. The color moments and Gabor filter features of the divided different meaningful regions are finally utilized to construct local feature vectors, which are used to index and represent the different part of the image to apply CBIR. Since a bank of filters at different scales and orientations can extract the texture frequency and orientation information effectively, the feature vector is computed by filtering the texture image with 6 orientations and 4 scales and computing the mean and standard deviation of the output in the frequency domain. Figure 3 shows some typical results of the individual step in the proposed three-level hierarchical structure for CBIR.

3 Experimental results and analysis

In order to verify the effectiveness of the proposed scheme, a huge number of retrieval experiments are used to evaluate the performance of our proposed CBIR scheme. In the retrieval experiment, sixty directional and salient patches are extracted in parallel, and the directional patches detection is based on 1-level wavelet decomposition, while the salient patches are extracted with wavelet transform of 3-level decomposition, both the two patches extraction are using db4 wavelet (which is an overlapping wavelet with a compact support, and the regularity p is 4). The DBSCAN clustering algorithm is performed to divide the salient patches into different salient regions/groups, and the value of α is empirically set to $\alpha = 0.02$. After the input parameter α is set, the neighborhood radius threshold *Eps* can be computed in a self-adaptation way according to Eq. (14); next, the density-based clustering algorithm can divide those directional and salient patches into different meaningful subclasses, as is shown in the bottom-level in Fig. 1, which then can be used to form bottom-level-local feature vectors to describe the different part of the image to achieve CBIR. Experiment results have shown that the above overall setting of the parameters can cluster both the directional and salient patches into different meaningful regions to reflect the local image characteristics, effectively. In addition, the proposed approach can efficiently avoid the over-segmentation problem usually emerged in traditional segmentation-based CBIR algorithms.

In the retrieval stage, a user can choose a query example/image, or selects an object or region of interest in an image using GUI (Graphical User Interface). Whereafter, wavelet transform with db4 wavelet is performed to the query image, and the hierarchical-local-feature extraction scheme works. As a result, as illustrated in Fig. 3c and e, the clustering algorithm can be performed to parallelly divide the extracted directional and salient patches into different regions, namely different meaningful sub-classes. Regarding the region specified in the query image using GUI, when there are multiple salient regions in the query image, the customer can choose a typical one or the whole image as input. Then, those images with feature vectors close to the feature vector of the specified region/image in the index space are returned. In our CBIR system, the hierarchical-local-feature extraction scheme can be realized by a tree structure of three levels (TREE Hierarchical data structure, in which a node with T child nodes expresses an image that are divided into different T visually meaningful regions in terms of their spatial distribution using the DBSCAN clustering algorithm), and the root node describes the global/holistic feature vector of the whole image. The main advantage of the developed hierarchical-local-feature extraction scheme is that it can take the global and hierarchical-local features into consideration simultaneously, and this scheme is suitable and very useful for network big-image-data indexing. In the similarity-distance-computing stage, the CBIR system directly outputs specified number of images with user-interested salient regions/objects of



(a) Original images.

(b) Perception-based directional patch extraction.

(c) Clustering the directional patches into regions to represent the local image properties.

(d) Improved salient patch extraction.

(e) Clustering the improved salient patches into regions to represent the local image properties.

Fig. 3 Some typical results of the separate steps in the proposed three-level hierarchical structure

whose indices are closest to the query feature vector, based on the weighted Euclidean distance [17]. The returned images are sorted in order that the candidate with the most resembling or similar salient region of user's interest searches in the first place. Even though an image contains some regions which are similar to the query key (feature vector), the searched image displays only once in the retrieval interface (Fig. 4).

We perform extensive experiments to evaluate the performance of the proposed approach. A hybrid and huge image dataset is downloaded from the internet network based on Corel

Fig. 4 Some examples used in the retrieval experiments

images Database [3], McGill Calibrated Color Image Database [29], SIMPLIcity Image Database [37], Ground Truth Database [12] and some images collected from the internet network by our research groups. The image dataset is comprised of more than 16,000 different categories of images. For CBIR experiments, the test set is randomly selected and includes 80 species, 4000 images in total. Figure 3 shows some examples, with various regions/objects in an image, used in our retrieval experiments.

For comparison, we also compare our developed method with seven state-of-the-art CBIR methods. The seven methods are:

(1) a classic CBIR method using wavelet-based salient points [42],
(2) three region-based image retrieval methods [9, 17, 44],
(3) a method using color difference histogram proposed by Liu and Yang [24],
(4) an fusion framework for color image retrieval proposed by Walia et al. [45], and
(5) a method using Quaternionic Local Ranking Binary Pattern proposed by Lan et al. [22].

In order to objectively compare the developed CBIR system with the other seven state-of-the-art CBIR approaches, the precision and recall, which are the two widely used evaluation measurements in CBIR, are calculated to assess the different schemes. Precision is defined as the fraction of the number of relevant images are retrieved over those of the number of all returned images, while recall assesses the ratio of the number of relevant image are retrieved over those of the number of the total relevant images, as follows:

$$\text{Preision} = \frac{\text{the number of relevant images are retrieved}}{\text{the number of all returned images}}, \tag{15}$$

and

$$\text{Recall} = \frac{\text{the number of relevant images are retrieved}}{\text{the number of the total relevant images}}. \tag{16}$$

Firstly, the retrieval performance are evaluated in terms of the average precision $P(n)$ in a intuitive way, which is defined as a function of the total number of correctly retrieved images in the first n returned images (Eq. (15)). In our experiments, the precision $P(n)$ for the first n ($n = 10$, 30, 50 and 70) retrieval results of seven state-of-the-art methods and our method are accounted. The new proposed method are compared with the other seven approaches are tabulated in Table 1, to assess the average precision. From Table 1, it can be seen that our method can achieve a superior average precision to the other seven state-of-the-art CBIR methods. As tabulated in Table 1, the average retrieval results show the effectiveness of the developed CBIR system.

Thank to the two evaluation measurements - precision and recall are mutually restricted in definition [17, 25], Precision-recall curve is also utilized to assess the overall image-retrieval performance. In practice, the average precision-recall curve (PR-curve) is produced by calculating different numbers of returned images n. Figure 5 shows the retrieval performance of the eight different methods, according to the average precision-recall curve. Obviously, as can be seen from Fig. 5, our proposed scheme of three-level hierarchical structure for content-based image retrieval produces a better comprehensive performance than the other state-of-the-art CBIR algorithms. That's because that when a input image contains local features, for instance, including some different meaningful regions or various objects, the CBIR systems using only global/holistic features will be degraded significantly [22, 45]. Our proposed hierarchical-local-feature extraction scheme, which comprises three levels and also considering both the global and local features simultaneously, utilizing local characteristics of images can remove some influence arose from these irrelevant texture areas or image background (see Fig. 3).

Furthermore, as the Fig. 5 illustrates, the proposed three-level hierarchical structure achieves much better retrieval performance than the previous algorithm proposed by Jian et al. [17]. This attributes to the flexibility of the deeply three-level hierarchical indexing structure, and the strong representation-ability of the directional and the improved salient patches for representing the local image properties. For the latter algorithm [17], which mainly focuses on extracting salient regions singly, is unable to capture the local feature of perception-based directional patches. In sum, the devised three-level hierarchical indexing structure is efficient and effective for network big-image-data representation and indexing, and the integrated feature vectors with the REE Hierarchical data structure are suitable for

Table 1 Average precision $P(n)$ of eight different methods

Average precision (%) Algorithms:	$P(10)$	$P(30)$	$P(50)$	$P(70)$
Jian et al. [17]	52.88	25.21	14.27	12.34
Fauqueur et al. [9]	47.33	19.79	11.40	9.25
Vieux et al. [44]	52.50	24.16	14.03	12.04
Tian et al. [42]	45.21	19.35	9.15	7.94
Liu et al. [24]	56.05	28.12	16.90	14.15
Walia et al. [45]	51.25	22.20	12.37	10.64
Lan et al. [22]	55.26	26.32	16.11	12.85
Our method	**58.19**	**29.74**	**17.48**	**14.82**

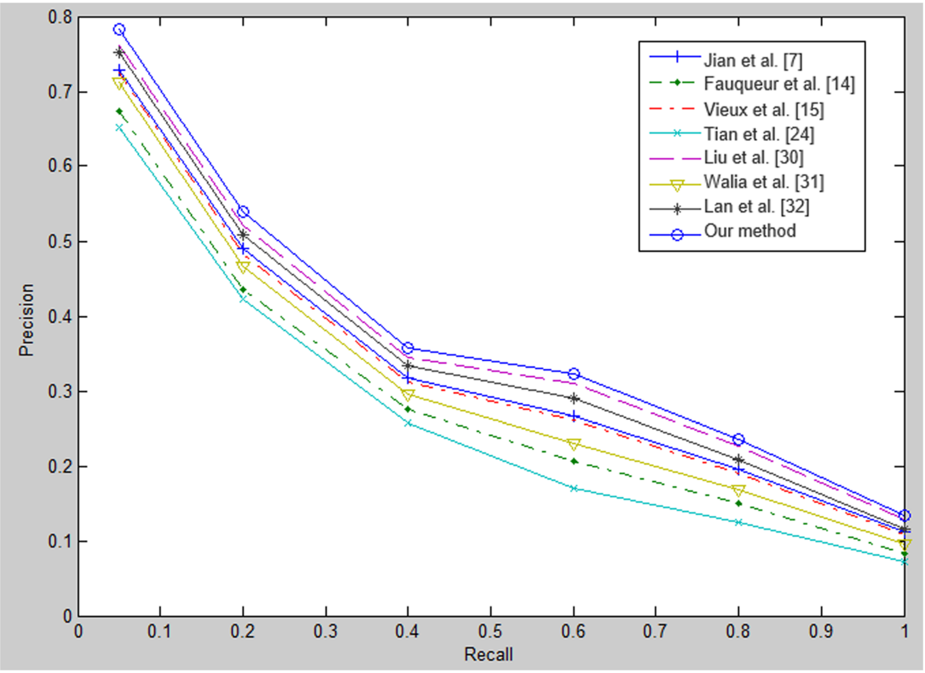


Fig. 5 Precision-recall curve of eight different methods

depicting an image with various objects or different meaningful regions. Experimental results show that the proposed hierarchical-local-feature extraction scheme is more effective than the other seven state-of-the art algorithms.

4 Conclusion and discussion

Since global features usually cannot capture various local characteristics of different parts in the image, in this paper, a hierarchical-local-feature extraction scheme with three-level TREE Hierarchical data structure is introduced to represent an image. Our purpose is not to accurately segment each single object in the image using traditional image-segmentation based methods, but to focus and extract these salient regions that are visually meaningful in view of human visual perception for CBIR. With the aid of wavelet transform, the directionality patches detector and the improved wavelet-transform salient patch detector based on Chauvenet's criterion are developed to describe the local image properties. Extensive experimental results verify that the developed CBIR scheme has produced promising retrieval performance.

In our future work, we hope to investigate and design a compact scale-rotate-shift-invariant representation of image, including edge representation matrix and corner representation matrix, to further improve image search results.

Acknowledgments This work was supported by National Natural Science Foundation of China (NSFC) (61601427, 61602229, 61771230); Natural Science Foundation of Shandong Province (ZR2015FQ011, ZR2016FM40); Shandong Provincial Key Research and Development Program of China (NO. 2017CXGC0701); Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

References

1. Alzu'bi A, Amira A, Ramzan N (2015) Semantic content-based image retrieval: a comprehensive study. J Vis Commun Image Represent 32:20–54
2. Carson C, Belongie S, Greenspan H (2002) Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans PAMI 24(8):1026–1038
3. Corel: Image Library University of California, Berkely. http://calphotos.berkeley.edu/use.html#download
4. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math 41: 909–996
5. ElAlami ME (2014) A new matching strategy for content based image retrieval system. Appl Soft Comput 14:407–418
6. Elleuch N, Ben Ammar A, Alimi AM (2015) A generic framework for semantic video indexing based on visual concepts/contexts detection. Multimed Tools Appl 74(4):1397–1421
7. Ester M, Kriegel HP, Sander J, Xu X (1996) A density based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, pp. 226–231
8. Fauqueur J, Boujemaa N (2002) Region-based retrieval: coarse segmentation with fine signature, IEEE ICIP, Rochester, NY, USA
9. Fauqueur J, Boujemaa N (2004) Region-based image retrieval: fast coarse segmentation and fine color description. J Vis Lang Comput 15:69–95
10. Gao L, Guo Z, Zhang H, Xu X, Shen HT (Sep. 2017) Video captioning with attention-based LSTM and semantic consistency. IEEE Trans Multimed 19(9):2045–2055
11. Gouet V, Boujemaa N (2001) Object-based queries using color points of interest. IEEE Workshop on Content-based Access of Image and Video Labraries, vol. 1, p 30–36
12. Ground Truth Database: Department of Computer Science and Engineering, University of Washington. http://www.cs.washington.edu/research/imagedatabase/groundtruth/_tars.for.download/
13. Jain AK, Farroknia F (1991) Unsupervised texture segmentation using Gabor filters. Pattern Recogn 24(12): 1167–1186
14. Jian MW, Dong JY (2007) Wavelet-Based Salient Regions and their Spatial Distribution for Image Retrieval, IEEE International Conference on Multimedia & Expo., p 2194–2197, 2–5 July
15. Jian M, Lam K-M (2014) Face-image retrieval based on singular values and potential-field representation. Signal Process 100:9–15
16. Jian M, Guo H, Liu L (2009) Texture classification using visual perceptual texture features and Gabor wavelet features. J Comput 4(8):763–770
17. Jian M, Dong J, Ma J (2011) Image retrieval using wavelet-based salient regions. Imaging Sci J 59(4):219–231
18. Jian M, Lam K-M, Dong J (2014) Facial-feature detection and localization based on a hierarchical scheme. Inf Sci 262:1–14
19. Jian M, Lam K-M, Dong J (2014) Illumination-insensitive texture discrimination based on illumination compensation and enhancement. Inf Sci 269:60–72
20. Jian M, Lam K-M, Dong J, Shen L (2015) Visual-patch- attention-aware saliency detection. IEEE Trans Cybern 45(8):1575–1586
21. Jiji GW, DuraiRaj PJ (2015) Content-based image retrieval techniques for the analysis of dermatological lesions using particle swarm optimization technique. Appl Soft Comput 30:650–662
22. Lan R, Zhou Y, Tang YY (2016) Quaternionic local ranking binary pattern: a local descriptor of color images. IEEE Trans Image Process 25(2):566–579
23. Lau HF, Levine MD (2002) Finding a small number of regions in an image using low-level features. Pattern Recogn 35(11):2323–2339
24. Liu G, Yang J (2013) Content-based image retrieval using color difference histogram. Pattern Recogn 46(1): 188–198
25. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282
26. Liu F, Zhang D, Shen L (2015) Study on novel curvature features for 3D fingerprint recognition. Neurocomputing 168(1):599–608

27. Long F, Zhang HJ, Feng DD (2003) Fundamentals of content-based image retrieval. In: Feng D, Siu WC, Zhang HJ (eds) Multimedia information retrieval and management-technological fundamentals and applications. Springer, Berlin
28. Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans PAMI 11(7):674–693
29. McGill Calibrated Colour Image Database http://tabby.vision.mcgill.ca/html/browsedownload.html
30. Muwei J, Dong J (2007) Combining color, texture and region with objects of user's interest for content-based image retrieval. Eighth ACIS International Conference on SNPD, p 713–718
31. Pavlidis T (2008) Limitations of content-based image retrieval. ICPR. http://www.theopavlidis.com/technology/CBIR/PaperB/vers3.htm
32. Sander J, Ester M, Kriegel HP et al (1998) Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min Knowl Disc 2(2):169–194
33. Sebe N, Lew MS (2003) Comparing salient points detectors. Pattern Recogn Lett 24(1–3):89–96
34. Sebe N, Tian Q, Loupias E, Lew MS, Huang TS (2001) Content-based Retrieval using Salient Point Techniques, IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01), Technical Demo, Electronic Proceedings, Kauai, Hawaii
35. Sebe N, Tian Q, Loupias E, Lew MS, Huang TS (2003) Evaluation of salient point techniques. J Image Vision Comput 21(13–14):1087–1095
36. Shen L, Bai L (2008) 3D Gabor wavelets for evaluating SPM normalization algorithm. Med Image Anal 12(3):375–383
37. SIMPLIcity Image Database: http://wang.ist.psu.edu/docs/related/
38. J. Song, Lianli Gao, Xiaofeng Zhu, Nicu Sebe (2017) Quantization based hashing: a general framework for scalable image and video retrieval. Pattern Recogn
39. Song J, He T, Gao L, Xu X, Shen H (2018) Deep region hashing for efficient large-scale instance search from images. AAAI
40. Sudhakar MS, Bagan KB (2014) An effective biomedical image retrieval framework in a fuzzy feature space employing phase congruency and GeoSOM. Appl Soft Comput 22:492–503
41. Taylor JR (1997) An introduction to error analysis, 2nd edn. University Science Books, Sausolito, California
42. Tian Q, Sebe N, Loupias E, Lew MS, Huang TS (2001) Image retrieval using wavelet-based salient points. J Electron Imaging 835–849
43. Tsai HH, Chang BM, Liou SH (2014) Rotation-invariant texture image retrieval using particle swarm optimization and support vector regression. Appl Soft Comput 17:127–139
44. Vieux R, Benois-Pineau J, Domenger J-P (2012) Content based image retrieval using bag-of-regions, 18th International Conference, MMM 2012, Klagenfurt, Austria, January 4–6, pp. 507–517
45. Walia E, Pal A (2014) Fusion framework for effective color image retrieval. J Vis Commun Image Represent 25(6):1335–1348
46. Wang Q, Yuan Y, Yan P, Li X (2013) Visual saliency by selective contrast. IEEE Trans Circ Syst Vid Technol 23(7):1150–1155
47. Wang Q, Yuan Y, Yan P, Li X (2013) Saliency detection by multiple-instance learning. IEEE Trans Cybern 43(2):660–672
48. Wang Q, Wan J, Yuan Y (2018) Locality constraint distance metric learning for traffic congestion detection. Pattern Recogn 75:272–281
49. Wang X, Gao L, Wang P, Sun X, Liu X (2018) Two-stream 3D convNet fusion for action recognition in videos with arbitrary size and length. IEEE Trans Multimed 20(3):634–644
50. Wang J, Zhang T, Song J, Sebe N, Shen H (2018) A survey on learning to hash. IEEE Trans Pattern Anal Mach Intell 40(4):769–790
51. Wang Q, Wan J, Yuan Y Deep metric learning for crowdedness regression. IEEE Trans. Circ Syst Vid Technol. https://doi.org/10.1109/TCSVT.2017.2703920
52. Xu Y-Y (2016) Multiple-instance learning based decision neural networks for image retrieval and classification. Neurocomputing 171:826–836
53. Yang M, Zhu P, Liu F, Shen L (2015) Joint representation and pattern learning for robust face recognition. Neurocomputing 168(30):70–80
54. Zhu Z, Jia S, He S, Sun Y, Ji Z, Shen L (2015) Three-dimensional Gabor feature extraction for hyperspectral imagery classification using a memetic framework. Inf Sci 298(1):274–287
55. Zhu Y, Jiang J, Han W, Ding Y, Tian Q (2017) Interpretation of users' feedback via swarmed particles for content-based image retrieval. Inf Sci 375:246–257

Muwei Jian received the PhD degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, in October 2014. He was a Lecturer with the Department of Computer Science and Technology, Ocean University of China, from 2015 to 2017. Currently, Dr. Jian is a Professor and Ph.D Supervisor at the School of Computer Science and Technology, Shandong University of Finance and Economics. His current research interests include human face recognition, image and video processing, machine learning and computer vision. Prof. Jian was actively involved in professional activities. He has been a member of the Program Committee and Special Session Chair of several international conferences, such as *SNPD 2007, ICIS 2008, APSIPA 2015, EEECS 2016, ICTAI2016, ICGIP 2016 and ICTAI 2017*. Dr. Jian has also served as a reviewer for several international SCI-indexed journals, including *IEEE Trans., Pattern Recognition, Information Sciences, Computers in Industry, Machine Vision and Applications, Machine Learning and Cybernetics, The Imaging Science Journal, and Multimedia Tools and Applications*. Prof. Jian holds 3 granted national patents and has published over 40 papers in refereed international leading journals/conferences such as *IEEE Trans. on Cybernetics, IEEE Trans. on Circuits and Systems for Video Technology, Pattern Recognition, Information Sciences, Signal Processing, ISCAS, ICME and ICIP*.

Yilong Yin received the Ph.D. degree from Jilin University, Changchun, China, in 2000. From 2000 to 2002, he was a Post-Doctoral Fellow with the Department of Electronics Science and Engineering, Nanjing University, Nanjing, China. He is currently the Director of the data Mining, Machine Learning, and their Applications Group and a Professor of the School of Computer Science and Technology, Shandong University, Jinan, China. His research interests include machine learning, data mining, and computational medicine.

Junyu Dong received his B.Sc. and M.Sc. in Applied Mathematics from the Ocean University of China (formerly called Ocean University of Qingdao) in 1993 and 1999, respectively. He won the Overseas Research Scholarship and James Watt Scholarship for his PhD study in 2000 and was awarded a Ph.D. degree in Image Processing in 2003 from the School of Mathematical and Computer Sciences, Heriot-Watt University, UK. Dr. Junyu Dong joined Ocean University of China in 2004. From 2004 to 2010, Dr. Junyu Dong was an associate professor at the Department of Computer Science and Technology. He became a Professor in 2010 and is currently the Head of the Department of Computer Science and Technology. Prof. Dong was actively involved in professional activities. He has been a member of the program committee of several international conferences, including the 4th International Workshop on Texture Analysis and Synthesis (associated with ICCV 2005), the 2006 British Machine Vision Conference (BMVC 2006) and the 3rd International Conference on Appearance (Predicting Perceptions 2012). Currently, Prof. Dong is the Chairman of Qingdao Young Computer Science and Engineering Forum (YOCSEF Qingdao). He is a member of ACM and IEEE. Prof. Dong's research interest includes texture perception and analysis, 3D reconstruction, video analysis and underwater image processing.

Kin-Man Lam received the Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Sydney, Australia, in August 1996. From 1990 to 1993, Prof. Lam was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in October 1996. He became an Associate Professor in 1999, and is now a Professor. Prof. Lam was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. In particular, he was the Secretary of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), the Technical Chair of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP 2004), a Technical Co-Chair of the 2005 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2005), a secretary of the 2010 International Conference on Image Processing (ICIP 2010), a Technical Co-Chair of 2010 Pacific-Rim Conference on Multimedia (PCM 2010), and a General Co-Chair of the 2012 IEEE International Conference on Signal Processing, Communications, & Computing (ICSPCC 2012), which was held in Hong Kong in August 2012. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008. Currently, he is the VP-Member Relations and Development of the Asia-Pacific Signal and Information Processing Association (APSIPA) and the Director-Student Services of the IEEE Signal Processing Society. Prof. Lam serves as an Associate Editor of IEEE Trans. on Image Processing, Digital Signal Processing, APSIPA Trans. on Signal and Information Processing, and EURASIP International Journal on Image and Video Processing. He is also an Editor of HKIE Transactions. His current research interests include human face recognition, image and video processing, and computer vision.