information

MDPI

Article

A Novel Local Structure Descriptor for Color Image Retrieval

Zhiyong Zeng

Faculty of Software, Fujian Normal University, Qishan Community, Fuzhou 350108, China; zzyong@fjnu.edu.cn; Tel.: +86-0591-22868470 (ext. 108)

Academic Editor: Willy Susilo
Received: 19 November 2015; Accepted: 15 February 2016; Published: 22 February 2016

Abstract: A novel local structure descriptor (LSD) for color image retrieval is proposed in this paper. Local structures are defined based on a similarity of edge orientation, and LSD is constructed using the underlying colors in local structures with similar edge direction. LSD can effectively combine color, texture and shape as a whole for image retrieval. LSH integrates the advantages of both statistical and structural texture description methods, and it possesses high indexing capability and low dimensionality. In addition, the proposed feature extraction algorithm does not need to train on a large scale training datasets, and it can extract local structure histogram based on LSD. The experimental results on the Corel image databases show that the descriptor has a better image retrieval performance than other descriptors.

Keywords: local structures' descriptor; local structures' histogram; edge orientation similarity; underlying color; color image retrieval

1. Introduction

Images are one of the popular media formats for communication and understanding of human society. With the rapid development of internet and multimedia techniques, ever-increasing images are available to the public. Therefore, people desire to get a more efficient image indexing tool. Image search has become one of the crucial issues in computer vision. Generally, people retrieve images in three ways. Text-based methods use keywords that are annotated on images to search image, which are widely used by many company's applications, such as Google and Baidu. However, this approach needs to label images manually, and manually annotating a large scale of images is time-consuming. Furthermore, the retrieval results may be inaccurate. Content-based image retrieval (CBIR) approach computes similarity between images by using low level features which describe the content of each image, and then the retrieved images have similarity in certain predefined threshold. However, low level features cannot describe image semantic concepts because of a huge semantic gap. In the last few years, many semantic-based image retrieval (SBIR) methods have been proposed, owing to the limits of relevant techniques, SBIR is an open problem so far [1,2]. To this day, CBIR is still one of the most effective image indexing methods.

As is well known, a process competing interactions among neurons will strengthen human visual attention, this means that a few elements of attention are selected and other irrelevant materials are suppressed by neurons [3]. There are close connections between visual features of each image and the human visual system, and the study on how to exploit visual attention mechanism for CBIR is a vital and challenging problem. To simulate visual processing procedure, we propose a novel local structure descriptor (LSD) for image retrieval in this paper. The novelty of this descriptor lies in: (1) due to sensitivity of human visual perception to the orientation and color, we use edge orientation and color

of each image to describe the local structure of each image, then we use local structure descriptor to simulate visual processing procedure. (2) The descriptor can not only describe image feature, but also effectively combines color, texture, shape and color layout as a whole simultaneously without any image segmentation, training and learning. (3) The dimensionality of LSD feature vector is low, its time complexity is linear, and LSD is easy to implement.

In the LSD, we initially focus on how to define local structures. We adopt the feature integration method of two-stage model to define local structure [4]. In the pre-attentive stage, we extract primitive features in special modules of feature maps. In the attentive stage, we require focal attention to integrate the separate features to form objects. Subsequently, we move our focus toward describing local structures. We use a local structure histogram (LSH) to extract feature vectors of each image in HSV (Hue, Saturation, Value) color space. The local structure descriptor can describe color, texture and shape of each image simultaneously. Therefore, the LSD combines advantages of both statistical and structural texture descriptors. Finally, we focus on improving the performance of LSD by studying the effects of different color space, parameter settings, gradient operators, and distance metrics. The experimental results of the large image database show that the LSD gets higher precision and recall than representative visual representation descriptors.

Our contributions are as follows:

(a) We design a new local structure descriptor to simulate human visual perception mechanism. The descriptor combines color, texture, shape and color layout as a whole. The dimensionality of its feature vector is low, which is very appropriate for large-scale image retrieval. The descriptor achieves better accuracy results on standard benchmarks than other descriptors.

(b) The detail experimental research we carried out adds our understanding of the effects of different color space, parameter settings, and gradient operators of the studied descriptor.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 describes the definition of LSD. The extraction of local structure histogram in HSV color space is presented in Section 4. In Section 5, similarity measurement is presented. The proposed descriptor is evaluated and compared with state-of-the-art descriptors on Corel image databases. The conclusion of this paper is presented in Section 7.

2. Related Works

Generally, researchers use different descriptors to represent color, texture and shape of image. They design different algorithms to extract image features and retrieve images. Because of simplicity and effectiveness, the color histogram has been extensively studied and used for image retrieval. Meanwhile, it is invariant to scale and orientation. However, it is difficult for color histogram to represent the image spatial structure. Instead, to exploit image spatial information, other color descriptors have been proposed [5,6]. Texture features present valuable information of smoothness, coarseness and regularity of objects such as fruits, skins, clouds, tree, *etc.*, texture-based schemes have been widely studied in CBIR [7]. Manjunath has designed MPEG-7 edge histogram descriptor (EHD) according to the spatial distribution of edge, which is a very effective texture descriptor for image representation [8]. In addition, the outlines of different objects in an image are usually different, and shape features have been widely used in CBIR for its capability in describing the object outline. The edge orientation autocorrelogram (EOAC) [9] is one of the most successful methods.

Local image features have attracted more and more attention in recent years. Various feature algorithms were presented in order to highlight different properties of an image such as gray and edge, *etc.* These methods are often based on distribution, and color features can be combined with texture features. Lowe *et al.* propose a so-called scale-invariant feature transform (SIFT) descriptor, which is an effective algorithm to detect and describe local feature of an image [10]. Wang *et al.* propose a novel retrieval scheme based on structure elements' descriptors (SEDs) [11]. SED effectively presents the spatial connection of color and texture by extracting the color and texture features of an

image. Liu *et al.* propose an image feature representing method called color difference histogram (CDH) [12]. CDH is completely different from the existing histogram techniques; it pays more attention to perceptually uniform color difference between color and edge orientation. Murala *et al.* utilize the orientation map to obtain one order derivative on horizontal and vertical directions based on Local Tetra Patterns (LTrPs) for texture image retrieval and improve retrieval performance [13]. Meng *et al.* compute the feature of an image and its spatial correlation by using salient points in the patch of an image, and they propose the learning algorithm of multi-instances, this method improves the average retrieval accuracy of image [14]. Wang *et al.* extract the color feature of an image by Zernike color distributed moments and compute texture feature by contourlet transform. Then, they combine color with texture for image retrieval, and the algorithm obtains better image retrieval performance [15]. Local descriptors are tolerant to varied illuminations, distortions, transformations and are robust to occlusion. Meanwhile, they can describe different characteristics of object appearance or shape without any image segmentation. However, it is still an important challenge how to develop computational models that describe color, texture, shape and spatial structure simultaneously.

In order to improve further image retrieval performance, many feature combination methods and relevant feedback techniques have been proposed recently [16–19]. Lee *et al.* propose a novel retrieval scheme that extracts the image feature by fusing Advanced Speed-Up Robust Feature (ASURF) and Dominant Color Descriptor (DCD). The system can run in real-time on iPhone and find a natural color image for mobile image retrieval [16]. Kafai *et al.* describe Discrete Cosine Transform (DCT) hashing for creating index structures for face descriptors, and this method can efficiently reduce the cost of the linear search and improve retrieval efficiency and accuracy [17]. Yang *et al.* propose a semi-supervised Local Regression and Global Alignment (LRGA) algorithm for data ranking and a semi-supervised long-term Relevance Feedback (RF) algorithm for using data distribution and the history RF information, and then they integrate the two algorithms into multimedia content analysis and retrieval framework [18]. Spyromitros-Xioufis *et al.* use the framework of a Vector of Local Aggregated Descriptors (VLAD) and Product Quantization to develop an enhanced framework for large-scale image retrieval, the system significantly improves the performance for image retrieval [19].

Until now, the latest large-scale image retrieval frameworks use the bag-of-words (BoW) descriptor [20]. In these methods, retrieval system extracts local features (usually SIFT [10]) from each image and assigns each feature to the nearest visual word from a visual vocabulary. The feature vector of BoW is high dimensional and sparse for each image. BoW-based methods achieve better the accuracy of image indexing, but they cannot generalize to more than millions images datasets on a single machine owing to high computational complex and memory limits. Recently, a few scalable algorithms have been developed in [21–23]. The feature vectors of these methods are more discriminative than BoW, and these feature vectors combine with powerful compression techniques.

One of the most successful algorithms is proposed in [21]. This method uses SIFT features instead of BoW with highly discriminative Fisher Vector [22] or its simpler variant—VLAD [23]. Exploiting these optimized vector representation, greatly better retrieval accuracies are achieved.

In recent years, deep learning has reported encouraging results on CBIR tasks. For example, Wan *et al.* use a framework of deep learning for CBIR, they find that deep learning can mimic the human brain that is organized in a deep architecture and processes information through multiple stages of transformation and representation [24]. Ng *et al.* extract convolution features from different layers of the deep convolutional networks and adopt VLAD encoding to encode features into a vector, they conclude that intermediate layers or higher layers with finer scales produce better results for image retrieval [25]. Zhang *et al.* combine deep convolutional neural network (CNN) with learning hash functions, deep CNN is utilized to train the model, discriminative image features and hash functions are simultaneously optimized [26]. Lin *et al.* propose an effective deep learning framework to learn binary hash codes by employing a hidden layer for representation of the latent concepts that dominate the class labels in a point-wised manner [27]. Deep learning has many advantages. One of them is it can learn hierarchical features for different semantic abstraction, which effectively improve

the performance of CBIR. However, deep learning needs some tricks in parameters tuning, and the feature dimensionality of the last layer is very high, which may affect its application in practice.

3. Local Structure Descriptors

Directly comparing image content is impractical for image search because the contents of image are significantly different. However, the structural information of the same class images often shows some similarity. We may regard that a semantic of a natural image is made up of many common local structures. If these local structures could be achieved and presented effectively, we could use them as a common base for different image matching. It is very important for image retrieval if shape is extracted along with color, texture and color layout of each image.

One key problem of the local structure descriptor is how to define local structure. Let $g(x, y)$ be a color image of size $M \times N$. Firstly, to detect local structure, we convert RGB color space into HSV color space, we quantize the color image into 72 colors in HSV color space, we use nine structure templates to detect edge orientation. The nine structure masks are nine 2×2 matrixes of different directions, which are displayed in Figure 1. Then, in the edge orientation map, we define the local structure. Thirdly, we construct the LSD using underlying color in local structures. Finally, we use local structure histogram to describe the LSD to represent image feature. The LSD not only could effectively represent each image, meanwhile, it could extract shape, texture, color and color layout feature of each image simultaneously. We present these steps in detail as follows:

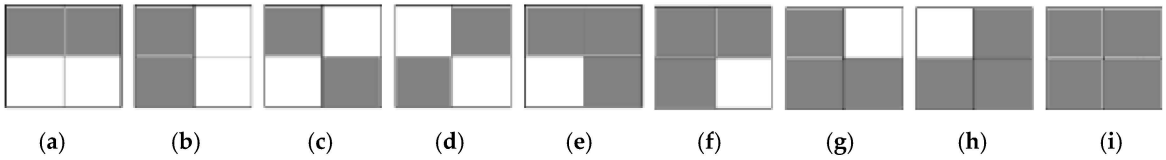


Figure 1. The definition of structure templates in different directions: (**a**) 0°; (**b**) 90°; (**c**) 45°; (**d**) 135°; (**e**) 225°; (**f**) 270°; (**g**) 315°; (**h**) 360°; (**i**) no direction.

3.1. Selection and Quantization of Color Space

HSV space is widely used for image feature extraction; it is made up of three components in terms of Hue (H), Saturation (S) and Value (V). HSV is usually modeled as a cylinder, the H component represents color type, its value is between 0–360°, with red at 0°, green at 120°, and blue at 240°. The S component describes the relative purity of color, its value is between 0–1. The V component represents the brightness of the color, its value is also between 0–1.

Human color perception could be imitated well by HSV color space l. In this paper, we select HSV color space, the color image is quantized into 72 bins. Specifically, we quantize the H, S and V components into 8, 3, 3 bins respectively. Hence, we obtain $8 \times 3 \times 3 = 72$ color combinations. Let $I(x, y)$ be the quantized color image, $I(x, y) = i, i \in \{0, 1, 2, ..., 71\}$.

3.2. Edge Direction Detection

Edge direction plays an important role in image recognition. The boundaries and texture structure of object is represented by the direction map of each image, the direction map provides a large amount of semantic concepts for the image. Hence, the detection of edge direction is a vital processing operation. Edge direction is detected by many existing edge detection operators such as Sobel operator, Canny operator, *etc.* However, these detectors can only detect gray level images and cannot detect color image because a color image has three color channels. In this subsection, the following methods are used to detect edge orientation for color image in HSV color space.

In Cartesian space, we define the dot product of vectors $a(x_1, y_1, z_1)$ and $b(x_2, y_2, z_2)$ as follows:

$$ab = x_1x_2 + y_1y_2 + z_1z_2 \tag{1}$$

so that

$$\cos\left(\overset{\wedge}{a,b}\right) = \frac{ab}{|a|\,|b|} = \frac{x_1x_2 + y_1y_2 + z_1z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2}\sqrt{x_2^2 + y_2^2 + z_2^2}} \tag{2}$$

Because cylinder coordinate system is used in HSV color space, to calculate the angle between vectors, HSV color space should be transformed into the Cartesian coordinate system. (H, S, V) represents a point in HSV color space, (H', S', V') represents the transformation of (H, S, V) in Cartesian space, $H' = S \cdot \cos(H)$, $S' = S \cdot \sin(H)$ and $V' = V$. The Sobel operator is applied to each of the H', S', V' channels of a color image in Cartesian space. We use two vectors $a(H'_x, S'_x, V'_x)$ and $b(H'_y, S'_y, V'_y)$ to denote the gradients along x and y direction, where the gradient in H' channel along horizontal direction is denoted by H'_x, and so on. Their dot product and norm can be defined as:

$$|a| = \sqrt{(H'_x)^2 + (S'_x)^2 + (V'_x)^2} \tag{3}$$

$$|b| = \sqrt{(H'_y)^2 + (S'_y)^2 + (V'_y)^2} \tag{4}$$

and

$$ab = H'_xH'_y + S'_xS'_y + V'_xV'_y \tag{5}$$

The angle between a and b is:

$$\cos\left(\overset{\wedge}{a,b}\right) = \frac{ab}{|a|\,|b|} \tag{6}$$

$$\theta = \arccos(\overset{\wedge}{a,b}) = \arccos\left[\frac{ab}{|a|\,|b|}\right] \tag{7}$$

After we calculate the edge direction of each pixel, we quantize the edge orientation into m *bins*, and $m \in \{6, 12, 18, 24, 30, 36\}$. We use $\theta(x, y)$ to represent the edge orientation map, as $\theta(x, y) = \phi$, $\phi \in \{0, 1, 2, ..., m-1\}$.

3.3. Definition and Extraction of Local Structure

The human visual system is sensitive to color and direction. Direction is a strong cue for topics described about an image. Obvious direction usually means definite pattern. However, there is no strong orientation and clear structure or specific pattern in many natural scenes. Although different contents are showed in the natural image, the common elements are included in these images. Different combination and spatial distribution of these fundamental elements lead to different local structures or patterns.

In order to find local structure with similar attributes such as edge direction and color, we use edge orientation map $\theta(x, y)$ to detect local structure, because edge direction is not sensitive to color and illumination variation, and it is robust to translation, scaling and small rotation. Let $\theta(x, y)$ be an edge direction map of size $M \times N$, we shift nine local structure templates shown in Figure 1 from top to bottom and left to right throughout orientation map to detect fundamental local structure respectively. In order to obtain a single local structure map of whole image, we use the steps described to detect as follows:

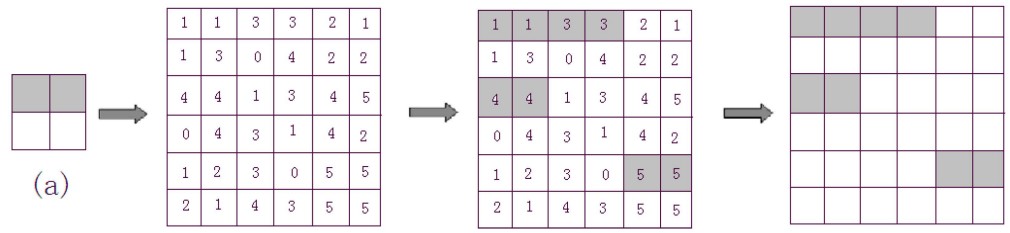
(1) Beginning from the point (0, 0), we shift 2×2 local structure template (a) from top-to-bottom and left-to-right throughout edge direction map $\theta(x, y)$ with a step length of two pixels along both vertical and horizontal directions. If the values of $\theta(x, y)$ in the corresponding structure template are equal, the values will be saved, otherwise, the values will be set zero. Then, we will obtain a local structure map $C_1(x, y)$.

(2) We use the other eight templates (b), (c), (d), (e), (f), (g), (h) and (i) throughout edge orientation map $\theta(x,y)$ to conduct the same operations as (1) step, respectively, we will obtain eight local structure maps $C_2(x,y)$, $C_3(x,y)$, $C_4(x,y)$, $C_5(x,y)$, $C_6(x,y)$, $C_7(x,y)$, $C_8(x,y)$, $C_9(x,y)$.

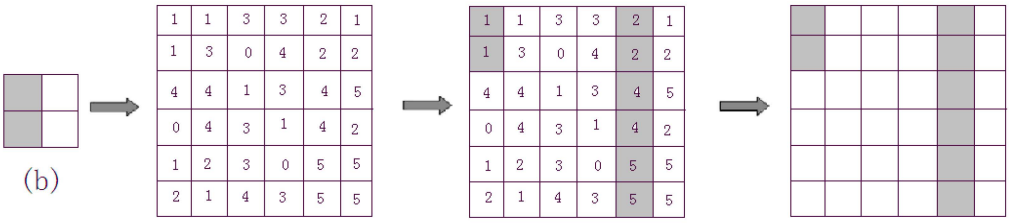
(3) Using C(x,y) to denote the final local structure map, C(x,y) is obtained by fusing nine local structure maps based on the following rules:

$$C(x,y) = \max\{C_1(x,y), C_2(x,y), C_3(x,y), C_4(x,y), C_5(x,y), C_6(x,y), C_7(x,y), C_8(x,y), C_9(x,y)\} \quad (8)$$

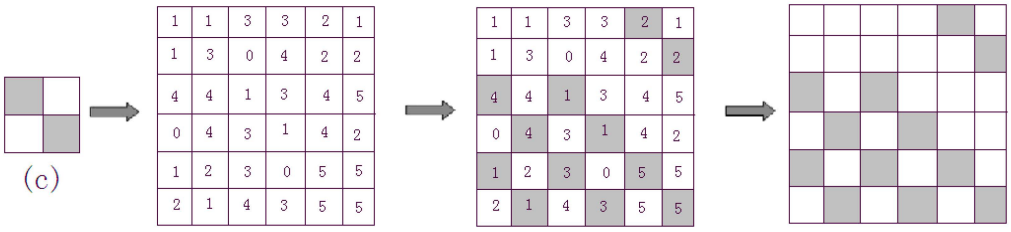
The extraction and fusing process of the above local structure map is illustrated in Figure 2.



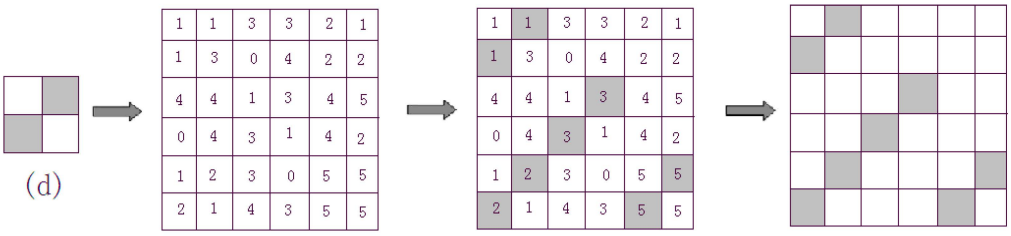
(**a**) 0° mask $\theta(x,y)$ local structure element localization $C_1(x,y)$.



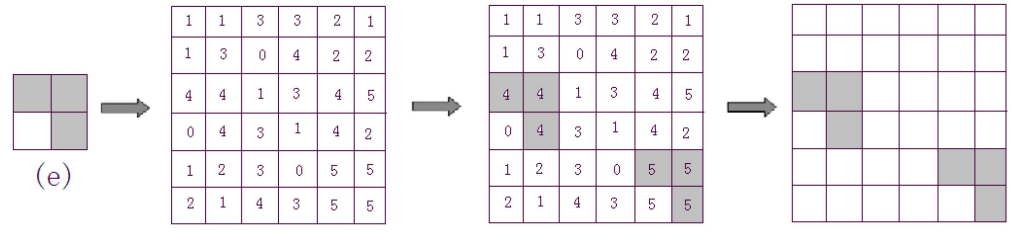
(**b**) 90° mask $\theta(x,y)$ local structure element localization $C_2(x,y)$.



(**c**) 45° mask $\theta(x,y)$ local structure element localization $C_3(x,y)$.

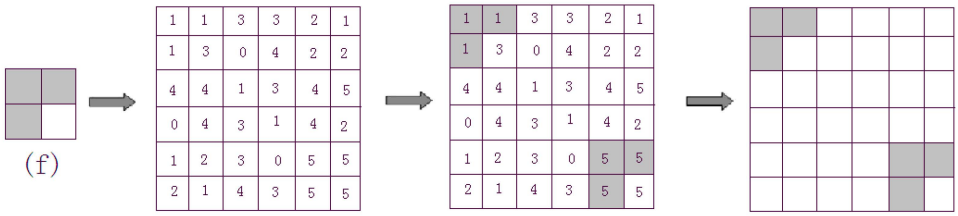


(**d**) 135° mask $\theta(x,y)$ local structure element localization $C_4(x,y)$.

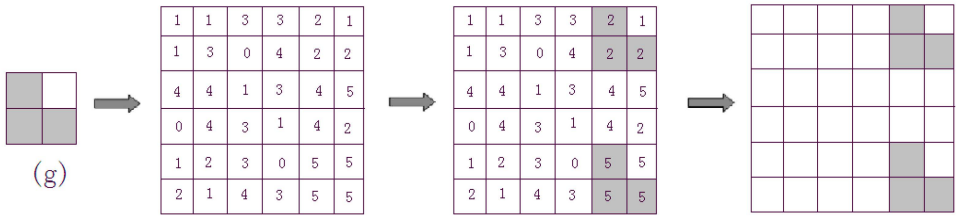


(**e**) 225° mask $\theta(x,y)$ local structure element localization $C_5(x,y)$.

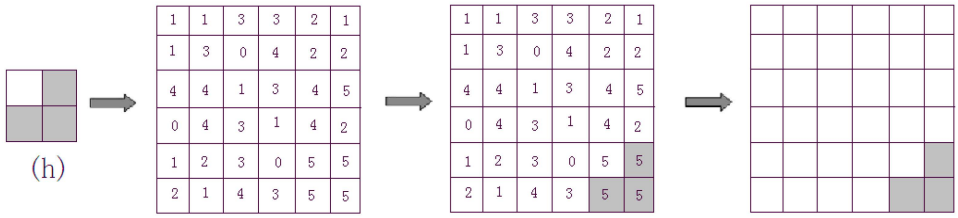
Figure 2. *Cont.*



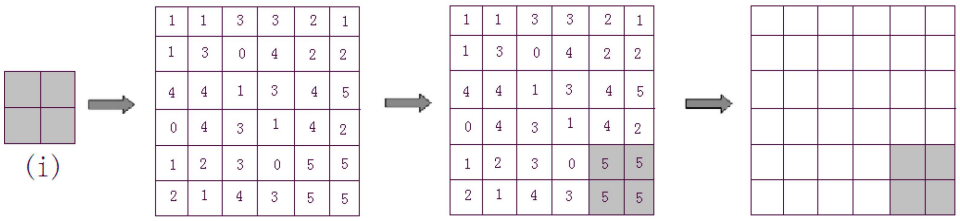
(**f**) 270° mask $\theta(x,y)$ local structure element localization $C_6(x,y)$.



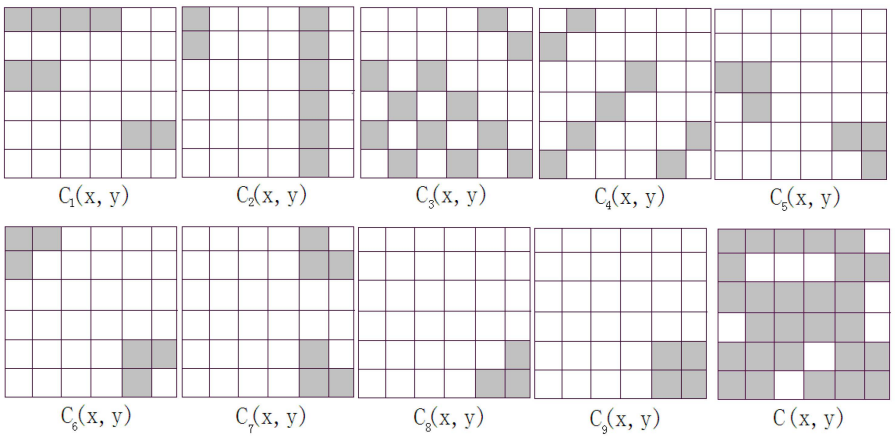
(**g**) 315° mask $\theta(x,y)$ local structure element localization $C_7(x,y)$.



(**h**) 360° mask $\theta(x,y)$ local structure element localization $C_8(x,y)$.



(**i**) No direction mask $\theta(x,y)$ local structure element localization $C_9(x,y)$.



(**j**)

Figure 2. Local structures' map extraction and fusion. (**a**) shows the local structure map extracted $C_1(x,y)$. Maps $C_2(x,y)$, $C_3(x,y)$, $C_4(x,y)$, $C_5(x,y)$, $C_6(x,y)$, $C_7(x,y)$, $C_8(x,y)$ and $C_9(x,y)$ can be extracted similarly in (**b**)–(**i**). (**j**) shows the fusion of nine Maps to form the final local structure map $C(x,y)$.

4. Feature Extraction

After the local structure map is extracted, the next step is to find a way to describe its feature. However, a critical problem is that we have to find a way that stimulates the visual system to some extent. Cortical visual system hypothesis is a framework for understanding cortical visual systems and visuomotor systems [4]. The framework is often called as the "what/where" and "what/how" pathways. It depicts two information processing streams which originate in the occipital cortex. The "what/where" pathways are responsible for color, shape, location and motion, respectively. The "how" pathway could be regarded as a generalization of the "where" pathway. The above analyses give an insight into how the relationship between image feature representation and visual attention mechanism is determined.

To present the content of image via imitating the visual attention mechanism to a certain extent, we consider the pixels with similar color in the local structure images as the stimuli in the visual field. We carry out the LSH method in accordance with the following rules: (1) what the local structure is; (2) where the local structure is; (3) how the local structure correlates with others. (1) and (2) are about the semantics of image based on perceptive space, and (3) is about the spatial representation of the local structures in a human brain.

After the local structure map $C(x,y)$ is extracted from the edge direction map $\theta(x, y)$, it is used as a mask to obtain the underlying colors from the quantized image $I(x,y)$. We will achieve the local structure map by reserving only the colors in the local structure image and other colors outside the image are set to empty. In the subsection, we apply $f(x, y)$ to denote the local structure image. The value of local structure map $f(x, y)$ can be denoted as $f(x, y) = i, i \in \{0, 1, 2, ..., 71\}$. We use local structure histogram LSH to extract feature vector of the image in HSV color space. In order to extract color, texture and shape feature of the $f(x, y)$ simultaneously, LSH will be calculated on every bin in HSV color space. LSH will be extracted by using two steps strategy described as follows:

(1) When $f(x, y) = i, i \in \{0, 1, 2, ..., 71\}$, count the number of LSD on nine LSD maps. In particular, when the local structure descriptor that denotes no direction has been counted, the other eight local structure descriptors should be counted again, because no direction means that every direction is possible.
(2) Calculate LSH based on the number of LSD.

The full color image is composed of local structure images. The attribute of the human visual system is fitted by local structure maps, because local structure images choose only a few points of attention and suppress the irrelevant material. Hence, the LSH not only depicts "what" and "how" colors and directions are utilized, but it also designates "where" and "how" the color and direction components are distributed within a certain range in the visual scenery. Hence, the different combination and spatial distribution of the local structures could be described by LSD, LSD could describes strong discriminative features such as texture, color, shape features and color layout information.

The process of extracting LSH is illustrated in Figure 3. Figure 3a displays the procedure of extracting LSH on an image when the value is 1, and the outcome of LSH is {1, 0, 1, 0, 0, 0, 0, 0, 0}. Figure 3b–e are the procedure of extracting LSH when the color values are 2, 3, 4, and 5. The final LSH of the image is {1, 0, 1, 0, 0, 0, 0, 0, 0, 2, 1, 0, 1, 0, 0, 1, 0, 0, 1, 2, 2, 1, 1, 1, 1, 2, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1}.

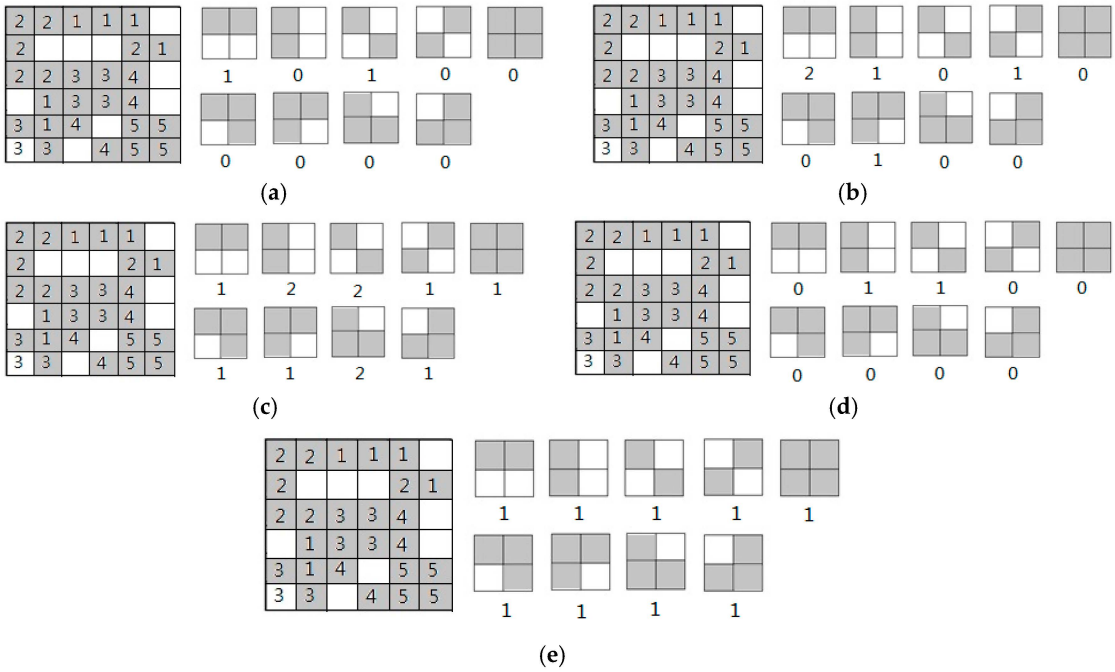


Figure 3. LSH extraction algorithm. (**a**)–(**e**) are the process of extracting LSH of nine bins, respectively. The number of which under the local structure is the number of the local structure of every bin.

In a natural image, LSH is extracted from 72 bins in HSV color space, all the nine local structure elements are extracted in each bin, respectively, so there are $72 \times 9 = 648$ bins extracted in HSV color space.

5. Similarity Measurement

For each image in a database, M bins feature vector $T = [T_1, T_2, ..., T_M]$ of the LSH is computed according to the above described method and stored in the database. Let $Q = [Q_1, Q_2, ..., Q_M]$ be M bins, the feature vector of a query image, the similarity measurement between a query image and any an image in the database can be defined as follows:

$$D(Q, T) = \sum_{i=1}^{M} |Q_i - T_i| \tag{9}$$

The above formula is called L_1 distance, which is the accumulative summary of two vectors difference, without square or square root operation on vectors. Hence, its time complexity is linear and is very suitable for large-scale image search.

6. Experiments and Results Analysis

6.1. Image Database

Two Corel image databases are used to test image retrieval performance in this paper, which are widely used for CBIR. The first one is Corel-1000 dataset, which includes 10 categories. There are 1000 natural images from diverse semantics such as scenes, horses, elephants, human, buses, flowers, buildings, mountains, foods and dinosaurs. The second one is Corel-10000 dataset, which includes 100 categories. There are 10,000 natural images from diverse semantics such as beaches, buses fishes and sunsets, *etc.* Experimental images contain different topics.

In following experiments, 10 images from each category in Corel-1000 dataset are randomly selected and used as query images. The precision and recall percentage for each category are calculated.

Then, average precision-recall pair percentage is calculated by the precision-recall pair percentage of 10 random images. In Corel-10000 dataset, we randomly choose 20 categories from 100 categories image. Then, 10 images are randomly selected from each category to use them as query images, and average precision and recall are computed.

6.2. Performance Measurements

We adopt Precision and Recall to evaluate the performance of the proposed method; these two metrics are the most commonly used metrics for evaluating image retrieval performance. Precision and Recall is defined as follows:

$$Precision = \frac{I_N}{N} \tag{10}$$

$$\text{Re}call = \frac{I_N}{M} \tag{11}$$

where I_N is the number of similar image retrieved, M is the total number of similar images in image database, and N is the total number of images retrieved. In the following experiments, to get the precision and recall values in Tables 1–3 *N and M* are set 12 and 100, respectively, for two datasets.

Table 1. The average retrieval performance of the local structure descriptor (LSD) under diverse color and direction quantization levels on Corel-1000 in RGB color space.

Color Quantization Level	Texture Orientation Quantization Level											
	Precision (%)						Recall (%)					
	6	12	18	24	30	36	6	12	18	24	30	36
128	93.43	93.32	93.13	92.82	92.91	93.12	9.40	9.38	9.36	9.32	9.33	9.35
64	92.50	92.42	91.86	91.97	91.99	92.25	9.28	9.27	9.21	9.22	9.22	9.25
32	89.85	90.06	89.50	89.41	89.64	89.66	8.96	8.97	8.90	8.91	8.94	8.93
16	82.12	82.63	82.38	82.55	82.22	82.35	8.02	8.12	8.09	8.11	8.04	8.05

Table 2. The average retrieval performance of the LSD under diverse color and direction quantization levels on Corel-1000 in HSV color space.

Color Quantization Level	Texture Orientation Quantization Level											
	Precision (%)						Recall (%)					
	6	12	18	24	30	36	6	12	18	24	30	36
192	99.05	98.91	98.73	98.44	98.69	98.56	10.31	10.30	10.28	10.25	10.27	10.26
128	99.09	98.57	98.85	98.73	98.60	98.49	10.33	10.26	10.28	10.28	10.25	10.24
108	98.88	98.54	98.46	98.72	98.72	98.16	10.18	10.12	10.12	10.16	10.15	10.09
72	98.20	98.24	98.30	98.45	98.54	98.12	10.06	10.05	10.08	10.09	10.14	10.05

Table 3. The retrieval results of LSD with different gradient operators for orientation detection.

Datasets	Performance	Gradient Operator				
		Proposed Operator	Sobel	Robert	LOG	Prewitt
Corel-1000	Precision (%)	98.20	97.83	97.15	96.21	97.52
	Recall (%)	10.06	9.96	9.98	9.81	9.95
Corel-10000	Precision (%)	52.26	51.85	51.54	51.18	51.62
	Recall (%)	5.87	5.75	5.73	5.68	5.74

6.3. Retrieval Results

In the following experiments, we adopt a different number of quantization level for texture and color to evaluate the retrieval performance in RGB and HSV color space. The RGB color space is

quantized to 16, 32, 64 and 128 bins, respectively, The HSV color space is quantized to 72, 108, 128 and 192 bins, respectively, and the texture orientation is quantized to 6, 12, 18, 24, 30 and 36 bins, respectively. As can be seen from the results in Tables 1 and 2 the proposed LSD has better performance in HSV color space than in RGB color space.

In order to test the results of the proposed direction detection operator, we use several classical edge detectors to detect gradient magnitude and direction, and the experimental results are listed in Table 3. We should note that the proposed operator works on the full color image, while the other four operators work on the gray level images. It can be seen from the retrieval results in Table 3 that the proposed direction operator achieves better results because it utilizes the color information that is ignored by the other detectors in direction detection.

We then validate the performance of the proposed distance metric and other popular distance metrics or similarity measurements. It can be seen from the retrieval results in Table 4 that our distance metric achieves better retrieval results than other distance metrics or similarity measurements. Meanwhile, comparing the results of L_1 distance and L_2 distance, we note that they have comparative performance with almost exactly precision and recall. However, L_1 distance is very simple to calculate with linear time requirement, unlike L_2 distance, it requires square or square root operations, which is costly for vector operation. Hence, our L_1 distance saves much computational cost and is very suitable for large-scale image database.

Table 4. The average results of LSD with different distance metrics.

Datasets	Performance	Distance or Similarity Measurement		
		L_1	Euclidian	Histogram Intersection
Corel-1000	Precision (%)	98.20	98.21	76.88
	Recall (%)	10.06	10.06	9.72
Corel-10000	Precision (%)	52.26	52.26	30.42
	Recall (%)	5.87	5.87	3.16

We compare the LSD representation with the state-of-the-art representations such as BoW [20], VLAD [22] and compressed Fisher kernel [23] on Corel-1000 and Corel-10000 datasets. These state-of-the-art representations are all parameterized by parameter k. Parameter k represents the number of centroids for BoW and VLAD, and to the number of mixture components in the Fisher kernel representation. We set k = 8000 for BoW and k = 64 for VLAD and compressed Fisher kernel, respectively. We use the same feature detection and description procedure as in [28] (Hessian–Affine extractor and the SIFT descriptor) since the code and the features are available online. We reduce the dimensionalities of these features from D = 8000 to D = 648 through PCA (Principal Component Analysis). We use the Flickr60k dataset to learn the PCA and the GMM (Gaussian Mixture Model) vocabularies. The Fisher vectors used in our experiments were computed by taking the gradient with respect to the mean parameters only. The average retrieval precision and recall curves of the above local descriptors are plotted in Figure 4. It can be seen from the retrieval results in Figure 4, LSD obtains better retrieval results than BoW [20], VLAD and Compressed Fisher kernel [22,23]. Though BoW is a good local feature representation, it ignores spatial information, BoW achieves the lowest retrieval results in the above descriptors. Fisher kernel and VLAD combine the strengths of generative and discriminative approaches, their feature representations are more discriminative and get better results than BoW. LSD imitates human visual mechanism to some extent, and it combines color, texture, shape and color layout as a whole. Meanwhile, LSD and its feature vector can be automatically extracted without any image segmentation, training and learning, they can be implemented simply, and the dimensionality of its feature vector is low, therefore, LSD is very suitable for large-scale image retrieval. The descriptor outperforms the previous best reported accuracy on standard benchmarks.

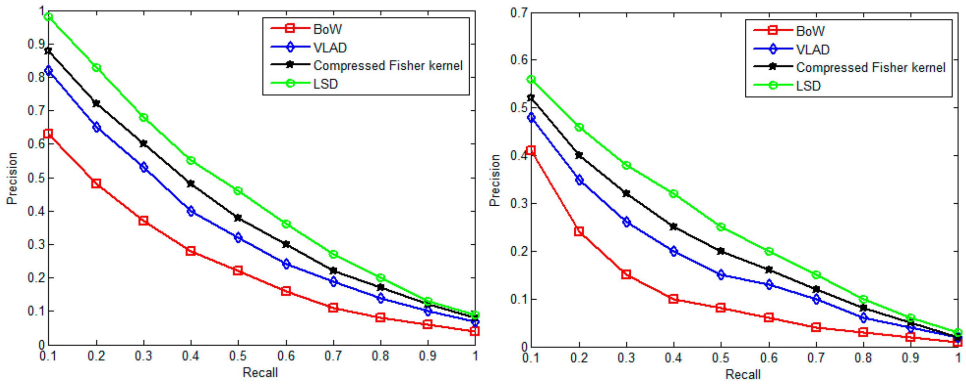


Figure 4. The comparison of average retrieval performance of four descriptors. (**a**) Corel-1000 datasets (**b**) Corel-10000 datasets.

Figures 5 and 6 show the retrieval examples on Corel-1000 and Corel-10000 datasets. In Figure 5, the query image is a bus image, it has clear shape feature and similar color. All the top 12 retrieved images belong to the bus category, which shows a good match of shape and color to query image. In Figure 6, the query image is a horse image, all of the top retrieved images are the horse category, which shows a good match of texture, shape and color to query the image.

Figure 5. Retrieval result for a bus.

Figure 6. Retrieval result for horses.

7. Conclusions

A novel local structure descriptor for color image retrieval is proposed in this paper. LSD can effectively describe the shape, texture, color, and color layout of each image without any image segmentation, learning and training. Hence, its implementation is easy and its time complexity is linear. LSD can be regarded as a generalized visual attribute descriptor because the descriptor imitates human visual mechanisms to some extent. The dimensionality of LSD feature vector is only 648, which is very suitable for large-scale image search. Our experimental results on two Corel image databases show that the proposed descriptor has strong discriminative capability for color, texture and shape features and outperforms the BoW, VLAD and Compressed Fisher kernel descriptors significantly.

Acknowledgments: This research is partially sponsored by the National Natural Science Foundation of China (No. 60805016), a major project of the Industrial Science and Technology of Fujian Province of China (No. 2013H0020).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Liu, Y.; Zhang, D.; Lu, G.; Ma, W.Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognit.* **2007**, *40*, 262–282. [CrossRef]
2. Zhang, D.S.; Islam, M.M.; Lu, G.J. A review on automatic image annotation techniques. *Pattern Recognit.* **2012**, *45*, 346–362. [CrossRef]
3. Desimone, R. Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. B* **1998**, *353*, 1245–1255. [CrossRef] [PubMed]
4. Reily, R.C.O. The what and how of prefrontal cortical organization. *Trends Neurosci.* **2010**, *33*, 355–361.
5. Huang, J.; Kumar, S.R.; Mitra, M. Image indexing using color correlograms. In Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997; pp. 762–768.
6. Quellec, G.; Lamard, M.; Cazuguel, G.; Cochener, B.; Roux, C. Fast wavelet-based image characterization for highly adaptive image retrieval. *IEEE Trans. Image Process.* **2012**, *21*, 1613–1623. [CrossRef] [PubMed]
7. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2007; pp. 395–398.

8. Manjunath, B.S.; Ohm, J.R.; Vasudevan, V.V.; Yamada, A. Color and texture descriptors. *IEEE Trans. Circuit Syst. Video Technol.* **2001**, *11*, 703–715. [CrossRef]
9. Mahmoudi, F.; Shanbehzadeh, J. Image retrieval based on shape similarity by edge orientation autocorrelogram. *Pattern Recognit.* **2003**, *36*, 1725–1736. [CrossRef]
10. Lowe, D.G. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Wang, X.Y.; Wang, Z.Y. A novel method retrieval based on structure elements' descriptor. *J. Vis. Commun. Image Represent.* **2013**, *24*, 63–74. [CrossRef]
12. Liu, G.H.; Yang, J.Y. Content-based image retrieval using color difference histogram. *Pattern Recognit.* **2013**, *46*, 188–198. [CrossRef]
13. Murala, S.; Maheshwari, R.P.; Balasubramanian, R. Local tetra patterns: a new feature descriptor for content-based image retrieval. *IEEE Trans. Image Process.* **2012**, *21*, 2874–2886. [CrossRef] [PubMed]
14. Meng, F.J.; Guo, B.L.; Wu, X.X. Localized Image Retrieval Based on Interest Points. *Proced. Eng.* **2012**, *29*, 3371–3375.
15. Wang, X.Y.; Yang, H.Y.; Li, D.M. A New Content-Based Image Retrieval Technique Using Color and Texture Information. *Comput. Electr. Eng.* **2013**, *39*, 746–761. [CrossRef]
16. Lee, Y.H.; Kim, Y. Efficient image retrieval using advanced SURF and DCD on mobile platform. *Multimed. Tools Appl.* **2015**, *74*, 2289–2299. [CrossRef]
17. Kafai, M.; Eshghi, K.; Bhanu, B. Discrete Cosine Transform Locality-Sensitive Hashes for Face Retrieval. *IEEE Trans. Multimed.* **2014**, *16*, 1090–1103. [CrossRef]
18. Yang, Y.; Nie, F.P.; Xu, D. A Multimedia Retrieval Framework based on Semi-Supervised Ranking and Relevance Feedback. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 723–742. [CrossRef] [PubMed]
19. Spyromitros-Xioufis, E.; Papadopoulos, S. A Comprehensive Study over VLAD and Product Quantization in Large-Scale Image Retrieval. *IEEE Trans. Multimed.* **2014**, *16*, 1713–1728. [CrossRef]
20. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1470–1477.
21. Jegou, H.; Perronnin, F.; Douze, M.; Sanchez, J.; Perez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1704–1716. [CrossRef] [PubMed]
22. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
23. Perronnin, F.; Liu, Y.; Sánchez, J.; Poirier, H. Large-scale image retrieval with compressed fisher vectors. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3384–3391.
24. Wan, J.; Wang, D.; Hoi, S.C.H.; Wu, P. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. *ACM Int. Conf. Multimed.* **2014**. [CrossRef]
25. Ng, J.Y.-H.; Yang, F.; Davis, L.S. Exploiting Local Features from Deep Networks for Image Retrieval. In Proceedings of the IEEE International Conference on Vision and Pattern Recognition, Deep Vision Workshop, Boston, MA, USA, 7–12 June 2015; pp. 53–61.
26. Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; Zhang, L. Bit-Scalable Deep Hashing with Regularized Similarity Learning for Image Retrieval and Person Re-identification. *IEEE Trans. Image Process.* **2015**, *24*, 4766–4779. [CrossRef] [PubMed]
27. Lin, K.; Yang, H.-F.; Hsiao, J.-H.; Chen, C.-S. Deep Learning of Binary Hash Codes for Fast Image Retrieval. In Proceedings of the IEEE International Conference on Vision and Pattern Recognition, Deep Vision Workshop, Boston, MA, USA, 7–12 June 2015; pp. 27–35.
28. Jegou, H.; Douze, M.; Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision—ECCV 2008*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5302, pp. 304–317.