



Image retrieval based on gradient-structures histogram

Bao-Hua Yuan^{1,2} · Guang-Hai Liu¹

Received: 15 February 2019 / Accepted: 3 December 2019 / Published online: 24 January 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

Color perception and orientation selection are very important mechanisms of the human brain that have close relationships with feature extraction and representation. However, extracting low-level features by mimicking these mechanisms remains challenging. To address this problem, we present the gradient-structures histogram as a novel method of content-based image retrieval (CBIR). Its main highlights are: (1) a novel and easy-to-calculate local structure detector, the gradient-structures, which simulates the orientation selection mechanism based on the opponent-color space and connects it with low-level features, (2) a novel discriminative representation method that describes color, intensity and orientation features. It is convenient, as it does not require weight coefficients for color, intensity and orientation. (3) The proposed representation method has the advantages of being histogram-based and having the power to discriminate spatial layout, color and edge cues. The proposed method provides efficient CBIR performance, as demonstrated by comparative experiments in which it significantly outperformed some state-of-the-art methods, including the Bow method, local binary pattern histogram, perceptual uniform descriptor, color volume histogram, color difference histogram, some improved LBP methods and the Tree2Vector method in terms of precision/recall and AUC metrics.

Keywords Image retrieval · Edge detection · Orientation selection · Gradient-structures histogram

1 Introduction

Searching is one of the most popular activities on the Internet. Image searching, or *retrieval*, has also become a very hot topic in academia. Current image retrieval technologies are predominantly based on global or local features (e.g., color, texture, edges, spatial information, key points and salient patches). People can use search engines to easily find images, videos and documents on the Internet, with Google, Bing, Yahoo and ASK being the most popular ones. A critical challenge for search engines is in how to extract and represent features from the vast amount of image data available.

Feature extraction and representation have close relationships with the color perception and orientation selection systems of the human visual system. For instance, as shown in the two scenes in Fig. 1, we pay more attention to regions with obvious orientations; those with bright colors and obvious orientations are more attractive. If we can extract the visual features of regions with obvious orientations, especially those that also have bright colors, we can improve discriminative representation and the retrieval performance of content-based image retrieval (CBIR) systems. In this paper, we investigate CBIR based on low-level visual features by mimicking the orientation selection and color perception mechanisms of the human visual system, which is a challenging problem.

In the 1980s, Hubel and Wiesel described simple cells as being linear with bar-shaped or edge-shaped receptive fields [1]. The visual cortex can be considered as a large collection of feature detectors which are tuned to edges and bars of various widths and orientations [1, 2]. Therefore, it is possible to extract image features by simulating the orientation selection mechanism according to edges and bars of various widths and orientations. Inspired by this concept, a novel image feature representation method,

✉ Guang-Hai Liu
liuguanghai009@163.com

¹ College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China

² Department of Computer Science and Technology, Taizhou Institute of Science and Technology, Nanjing University of Science and Technology, Taizhou 225300, China



Fig. 1 Examples of images with regions of obvious orientation (right) and without them (left). We pay greater attention to obviously oriented regions, especially those with bright colors (color figure online)

namely the gradient-structures histogram (GSH), which is based on the detection of gradient-structures, is proposed for CBIR.

The highlights of this paper include: First, a novel local structure detector, namely gradient-structures, is proposed to detect local structures by simulating the human orientation selection mechanism based on an opponent-color space. It is very easy to calculate and can perform a connective function between the orientation selection mechanism and low-level features. Second, a novel discriminative representation method (GSH) is proposed to describe image contents using colors, edge orientations and intensities as constraints. Obvious gradient-structures may make greater contributions to the retrieval results; therefore, it is not necessary to set weight coefficients for color, intensity and orientation. Third, the proposed representation method has the advantages of a histogram-based method with discriminative power based on spatial layout, color and edge cues. Therefore, the GSH method can provide efficient CBIR performance.

The remainder of this paper is organized as follows: In Sect. 2, image retrieval techniques are introduced. The proposed descriptor is presented in Sect. 3. In Sect. 4, performance comparisons are made between three benchmark datasets. Section 5 concludes the paper.

2 Related works

For more than a decade, image retrieval has been a hot topic in the field of artificial intelligence. In this subsection, we focus on image representation for content-based and object-based image retrieval.

Global features and local features are two types of visual features used in current image retrieval techniques. Global features usually comprise color, texture and shape features. In the MPEG-7 standard, color, texture and shape feature descriptors are used for image retrieval [3–14, 42]. Many studies show that a combination of multiple visual features can improve discriminative power. Some algorithms have validated this [9–14], such as the methods based on LBP [6–8, 15, 43, 44], texton-based methods [9, 10, 13] and

methods that use co-occurrence histograms and color differences for image representation [9–12]. Shape features are commonly used for image retrieval [3], object recognition and shape analysis [16–19]. In many cases, image segmentation is needed for the extraction of shape features, which remains a difficult problem in image processing. In some cases, well-defined object boundaries can be detected to determine shapes. In [14], a saliency structure histogram (SSH) is proposed for content-based image retrieval, which has good discriminative power for color, texture, edge and spatial layout features. In order to avoid image segmentation, many researchers have used local features (e.g., key points and salient patches) to extract image features [20–25].

Recently, dictionary learning methods have been reported in the literature for the use in object-based image retrieval, object recognition and scene categorization [20–32]. The bag-of-visual-words model, based on scale-invariant feature transform (SIFT), has been widely used in large-scale, object-based image retrieval applications. The standard Bow baseline can be considered as a state-of-the-art method [26]. Based on the bag-of-visual-words model, various methods have been proposed to represent image features. Wang et al. [28] proposed merging of the visual words into a large-sized initial codebook by maximally preserving class separability. Lobel et al. [29] proposed a new approach to visual recognition which jointly learns a shared, discriminative and compact mid-level representation and a compact, high-level representation using visual words. Liu et al. [30] proposed a single probabilistic framework to unify the merging criteria of visual words, which can be used for the creation of a compact codebook. Zhou et al. [31] proposed the binary SIFT (BSIFT) method for larger-scale, object-based image searches. Takahashi and Kurita [32] proposed a mixture of subspace image representation and compact coding for the large-scale, object-based image retrieval. Zhang et al. [45] proposed the learning of a vectoral representation of tree-structured data for CBIR and proposed a learning-based framework for video content-based advertising [46].

Besides the Bow-based methods, deep learning techniques represented by CNNs have been shown to be effective in various vision-related tasks, including classification and image retrieval, in recent years [33, 42]. However, Bow- and CNN-based methods have very high computational requirements compared to traditional global feature-based methods, such as low-level feature-based histograms. Thus, low-level feature representation within the CBIR framework needs to be studied further. The scope of the present study is limited to the use of low-level features in image representation.

3 The gradient-structures histogram

As mentioned before, it is possible to extract image features by simulating the human orientation selection mechanism based on the perception of edges and bars of various widths and orientations. Here, we propose a detector and discriminative representation system within the CBIR framework.

The detector is defined as a gradient-structures detector, which extracts edges and bars of certain widths and various orientations by simulating an orientation selection mechanism in an opponent-color space. Basing on the gradient-structures detector, we propose a discriminative representation system, namely the gradient-structures histogram (GSH), to describe image content and use it for CBIR. A flow diagram of the proposed detector and representation system within the CBIR framework is illustrated in Fig. 2.

3.1 Color quantization

Color plays an important role in visual perception and saliency detection [34, 35]. The HSV color space can mimic human color perception well and can be represented as a cylinder [5, 11–14, 34, 35]. This allows color information to be extracted easily. Color processing begins at a very early stage in the visual system (even within the retina) through initial opponent-color mechanisms [36]. In the gradient-structures model, both the HSV and opponent-color spaces are adopted to extract visual features. The opponent-color space is used to extract edge features, whereas the HSV color space is used to extract color, intensity and orientation features using feature quantization.

Let R, G and B be the red, green and blue components of a color image and then normalize them into [0,1] so that the yellow (Y) component can be defined as $Y = (R + G)/2$. These values can be transformed into values of the three

opponent-color channels: red versus green (RG), blue versus yellow (BY) and white versus black (WB). They can be defined as $RG = R - G$, $BY = B - Y$ and $WB = (B + R + G)/3$.

In order to maintain the visual appearance of the original image, a reduction in the number of colors is implemented. This implementation, called *color quantization*, is an important technique in CBIR using color features [9–14]. The concept of color quantization can be extended to extract other visual features, such as brightness or intensity quantization, orientation quantization and edge quantization. In order to describe image contents, the color, intensity and edge orientations are quantized using the quantization technique in the HSV color space [5, 13, 14]. The results of quantization are a color map, an intensity map and an edge orientation map. In color quantization, H , S and V are uniformly quantized into 6, 3 and 3 bins, respectively, resulting in a total of $6 \times 3 \times 3 = 54$ color bins. The results of color quantization are denoted as $M_C(x, y)$, and $M_C(x, y) = w, w \in \{0, 1, \dots, N_C - 1\}$, where $N_C = 54$.

Intensity information comes from the value component $V(x, y)$. After uniform quantization, an intensity map can be obtained and is denoted as $M_I(x, y)$, so that $M_I(x, y) = s$ and $s \in \{0, 1, \dots, N_I - 1\}$, where $N_I = 16$.

It is easy to calculate the edge of images using Sobel operators. According to Sobel operators and the value component $V(x, y)$, we can obtain the edge orientation $O(x, y)$. After uniform quantization, an edge orientation map can be obtained and is denoted as $M_O(x, y)$, and $M_O(x, y) = v, v \in \{0, 1, \dots, N_O - 1\}$, where $N_O = 60$.

3.2 Gradient-structures detection

Visual information processing in the visual system starts with opponent-color mechanisms [34]. Edge features are

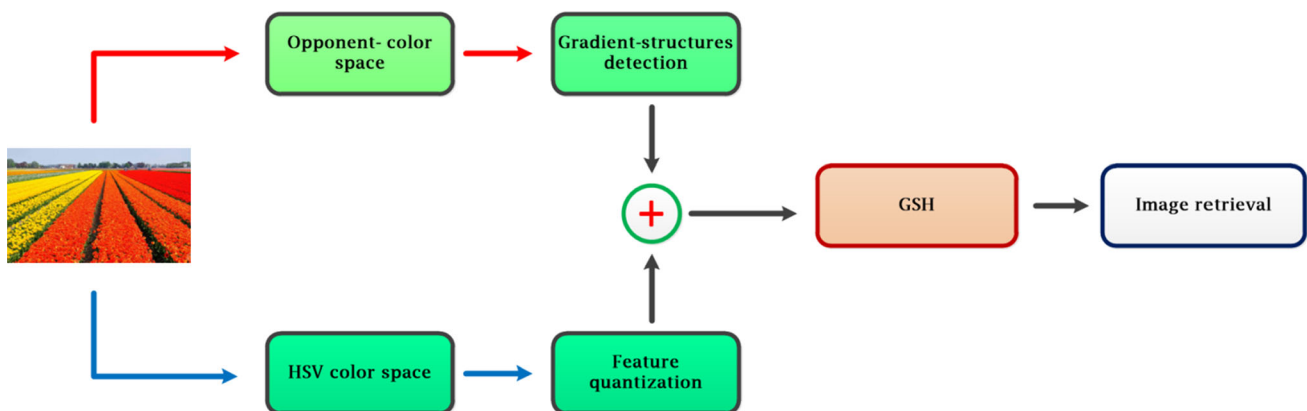


Fig. 2 Flow diagram of the proposed detector and discriminative representation system within the CBIR framework. Red arrows denote the procedure for simulating the orientation selection

mechanism. Blue arrows denote the procedure for feature quantization using color, orientation and intensity in the HSV color space (color figure online)

very important to visual perception [2, 36, 37]. In this paper, opponent colors are adopted for edge detection.

In our prior studies [9–12], Sobel operators were utilized to detect edges. Here, we still use this method to detect edges along the *RG*, *BY* and *WB* components in the opponent-color system. We use Sobel operators because they are insensitive to noise and have low computational requirements [37]. The values of *RG*, *BY* and *WB* are normalized into [0, 1], and their edge maps are denoted as $g_{RG}(x, y)$, $g_{BY}(x, y)$ and $g_{WB}(x, y)$, respectively. An example of color edge detection with opponent-color components is shown in Fig. 3.

Components $g_{RG}(x, y)$, $g_{BY}(x, y)$ and $g_{WB}(x, y)$ are all uniformly quantized into N_g bins denoted as $G_{RG}(x, y)$, $G_{BY}(x, y)$ and $G_{WB}(x, y)$, respectively. So, in this paper, $g_{RG}(x, y) = w, w \in \{0, 1, \dots, N_g - 1\}$, $g_{BY}(x, y) = w, w \in \{0, 1, \dots, N_g - 1\}$, and $g_{WB}(x, y) = w, w \in \{0, 1, \dots, N_g - 1\}$, where $N_g = 16$. In order to simulate the orientation selection mechanisms of the human visual system, $G_{RG}(x, y)$, $G_{BY}(x, y)$ and $G_{WB}(x, y)$ are further processed and used to detect gradient structures.

The gradient-structures are obtained from the consistency of edges. The term *consistency* is defined as when edges have the same gradient value in the same direction, let there be a 3×3 block in $G_{RG}(x, y)$ and (x, y) be a discrete coordinate. The values of the center coordinates (x_0, y_0) of the block are denoted as $g(x_0, y_0)$. Let there be two coordinates, (x_1, y_1) and (x_2, y_2) , on both sides of the central coordinate (x_0, y_0) , respectively. If $g(x_1, y_1) = g(x_0, y_0) = g(x_2, y_2)$, we consider such a local structure as having consistency of edge. In this case, the angle between the local structure and horizontal direction is denoted as a , $a = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$, which also denotes the sense of direction.

In the processing of gradient-structures detection, the most important factor is that the direction and gradient have the same values. As shown in Fig. 4c, if a gradient structure is found, its gradient values are kept and the gradient values of the remaining pixels within the 3×3 block are set to values of $N_g - 1$. Otherwise, all gradient

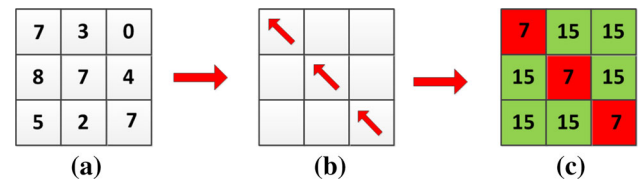


Fig. 4 Gradient-structures detection: **a** a 3×3 block in the quantized edge map $g(x, y)$, **b** the consistency of gradient detection in a 3×3 block of $g(x, y)$, and **c** in a gradient structure, we keep the original gradient values in pixels which have consistency of orientation, and the remaining pixels are set to values of $N_g - 1 = 15$; in this case, the angle $a = 135^\circ$

values of the 3×3 block are set to $N_g - 1$ values. By moving the 3×3 block from left-to-right and top-to-bottom throughout the input image using a pixel as the interval, we can obtain a map of the gradient-structures, which is denoted as $S_{RG}(x, y)$.

Using the same implementation in $G_{BY}(x, y)$ and $G_{WB}(x, y)$, $S_{BY}(x, y)$ and $S_{WB}(x, y)$ are also obtained. In order to integrate edge features according to feature integration theory [38], we calculate the final edge map in the manner of winner-take-all.

$$S(x, y) = \max(S_{RG}(x, y), S_{BY}(x, y), S_{WB}(x, y)) \quad (1)$$

In order to enhance the gradient-structures, the Sobel operators are utilized to detect edges in $S(x, y)$ again and denote the edge map as $E_s(x, y)$. It is obvious that the gradient-structures are very easy to calculate.

3.3 Feature representations

In global feature representation, local structure and spatial frequency play important roles in addition to low-level features (e.g., intensity, color and edge orientation); this is especially true in histogram-based image representation. Here, we consider local structures from a different perspective, which focuses on consistency in edge orientation. A novel image feature representation method, namely the gradient-structures histogram (GSH), is proposed for CBIR.

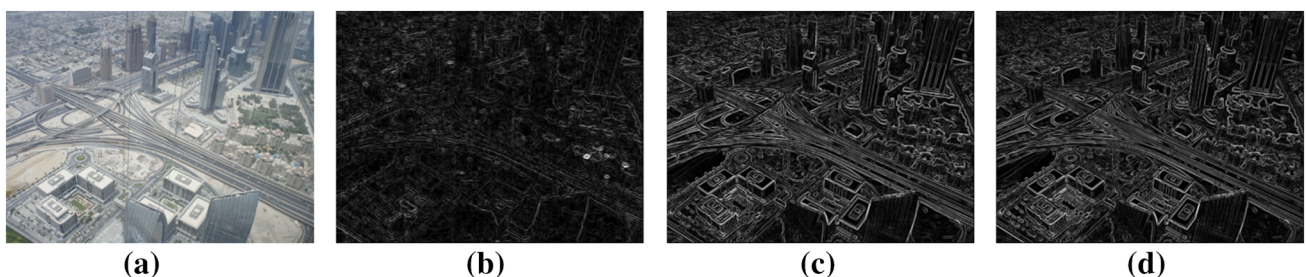


Fig. 3 Example of color edge detection with opponent-color components: **a** original color image; **b** edge magnitudes obtained from the RG component; **c** edge magnitudes obtained from the BY component; and **d** edge magnitudes obtained from the WB component

Let there be two pixel locations $(x - \Delta x, y - \Delta y)$ and $(x + \Delta x, y + \Delta y)$, on each side of the central pixel locations (x, y) in an image, where Δx and Δy are the offsets of the x -axis and y -axis, respectively. The gradient-structures histograms, using color, edge orientation and intensity as constraints, are denoted as H_C , H_O and H_I , respectively.

Their formulae are as follows:

$$H_C(M_C(x, y)) = \sum \sum \left(1.0 - \frac{1}{\pi(1 + E_s(x, y)^2)} \right) \quad (2)$$

where

$$M_C(x - \Delta x, y - \Delta y) = M_C(x, y) = M_C(x + \Delta x, y + \Delta y).$$

$$H_O(M_O(x, y)) = \sum \sum \left(1.0 - \frac{1}{\pi(1 + E_s(x, y)^2)} \right) \quad (3)$$

where

$$M_O(x - \Delta x, y - \Delta y) = M_O(x, y) = M_O(x + \Delta x, y + \Delta y).$$

$$H_I(M_I(x, y)) = \sum \sum \left(1.0 - \frac{1}{\pi(1 + E_s(x, y)^2)} \right) \quad (4)$$

where

$$M_I(x - \Delta x, y - \Delta y) = M_I(x, y) = M_I(x + \Delta x, y + \Delta y).$$

In order to integrate the primary visual features (i.e., intensity, color and edge orientation) and spatial frequency information into a single whole unit, the gradient-structures histogram of a full color image is defined as

$$H = \text{conca}\{H_C, H_O, H_I\} \quad (5)$$

In Formula (5), $\text{conca}\{.\}$ denotes the concatenation of H_C , H_O and H_I . In the concatenation, color, orientation and intensity have the same weight.

Using color, edge orientation and intensity as constraints, obvious the gradient-structures may make a greater contribution to the retrieval results; therefore, it is not necessary to set weight coefficients for color, intensity and orientation. For example, if gradient-structures of color are obvious, the retrieval results may show a good match for color features but not other types. The vector dimensions of the gradient-structures histogram are $54 + 60 + 16 = 130$ bins.

4 Experimental results

After image representation, the comprised algorithms, distance metrics and datasets must be selected for performance comparisons. The gradient-structures histogram (GSH), color volume histogram (CVH) [5], local binary pattern (LBP) histogram [15], color difference histogram

(CDH) [12], Bow histogram (SIFT-based) [26] and perceptual uniform descriptor (PUD) [41] were selected for comparison. CDH and CVH are two previous techniques we developed for CBIR, and the Bow histogram is considered a state-of-the-art method for object retrieval and recognition.

In the experiments, two image subsets were sampled for use as query images. These consisted of 10% of the total number of images in each dataset. The performance was evaluated using the average results of each query in terms of precision and recall. The codebook size for Bow was set as $k = 1000$, and the cosine metric was used as the baseline of the Bow method. In the PUD method, the Euclidean distance is adopted as the similarity measure. An LBP histogram with a dimensional feature vector of 256 bins, and using the average values of an LBP histogram based on R , G and B components. The L1 distance was adopted as the LBP histogram similarity measure.

For further comparison with some improved LBP methods and some other methods which work on tree-structured data, including local texture pattern (LTP) [43], local tetra patterns (LTrPs) [44], learning a vectorial representation for tree-structured data (Tree2Vector) [45] and the multilayer SOM (MLSOM) [48], we adopted the area under the precision–recall curve (AUC) and the precision and recall of various numbers of retrieved images as the performance metrics and used the Corel-1000 dataset as a benchmark dataset. All images in this dataset were used as query images.

4.1 Datasets

There are many datasets used in object-based image retrieval and object recognition. The Corel image dataset is the most commonly used dataset for testing the performance of CBIR. Images collected from the Internet can also serve as another data source for CBIR comparisons, especially for the retrieval of similar images. In this paper, the Corel dataset and web image collections were used for CBIR. The first dataset was the Corel-10k dataset, which contains 10,000 images in 100 categories, such as food, cars, sunsets, mountains, beaches, buildings, horses, fish, doors and flowers. Each category contains 100 images sized 192×128 or 128×192 pixels in JPEG format.

The second dataset used was the GHIM-10K dataset, which contains 10,000 images. All images in this dataset were obtained from the Internet or were created by the author (Guang-Hai Liu). It contains 20 categories, such as buildings, sunsets, fish, flowers, cars, mountains and tigers. Each category contains 500 images sized 400×300 or 300×400 pixels in JPEG format. Both the Corel-10K and GHIM-10K datasets can be downloaded from www.ci.gxnu.edu.cn/cbir.

For a fair comparison based on various metrics between improved LBP methods and some other methods that work with tree-structured data, the Corel-1000 dataset [47] was adopted as the third dataset. There are 1000 color images in ten categories, each category containing 100 images.

4.2 Distance metric

After we have extracted the image features from three datasets using the GSH, we store them in an SQL server 2008 database. T and Q are the feature vectors of a template image and a query image, respectively; T is an M -dimensional feature vector, $T = \{T_1, T_2, \dots, T_M\}$, and Q is also an M -dimensional feature vector, $Q = \{Q_1, Q_2, \dots, Q_M\}$. The Canberra distance between them is simply calculated as in [39], with a minor modification as follows:

$$\begin{cases} D(T, Q) = \sum_{i=1}^M w_i \cdot |T_i - Q_i| \\ w_i = \frac{1}{1.0 + |T_i| + |Q_i|} \end{cases} \quad (6)$$

$L1$ can be considered the distance, with w_i being the weight. In order to prevent the denominator being zero, a constant of 1.0 is added. This addition is a small modification to the Canberra distance.

In this paper, we set $M = 130$ bins for the proposed GSH in the three experimental datasets. The class label of the template image which yields the smallest distance will be assigned to the query image.

4.3 Performance measures

Using a specific type of performance metric is a very important issue in the comparison of CBIR experiments. In order to evaluate the effectiveness of the proposed algorithm, precision and recall metrics were adopted. In the field of image retrieval, they are the most common measurements used for evaluating performance [39]. *Precision* is the ratio of retrieved images relevant to the query. The comparisons made using the Corel-10K and GHIM-10K datasets were evaluated above a given cutoff point, considering only the top $N = 12$ positions. *Recall* is the ratio of images relevant to the query that are successfully retrieved [40].

$$\text{Precision} = \frac{|\{\text{relevant images}\} \cap \{\text{retrieved images}\}|}{|\{\text{retrieved images}\}|} \quad (7)$$

$$\text{Recall} = \frac{|\{\text{relevant images}\} \cap \{\text{retrieved images}\}|}{|\{\text{relevant images}\}|} \quad (8)$$

In this study, the number of relevant images was 100, which is also the total number of images in the database that are similar to the query.

The area under the precision–recall curve (AUC) is a single-number summary of the information in the precision–recall curve, which is related to both the precision and recall metrics. It can be defined as:

$$\text{AUC} = \sum_{N=2}^{N_{\max}} \frac{(P(N) + P(N-1)) \times (R(N) - R(N-1))}{2} \quad (9)$$

In Eq. (9), P and R denote precision and recall, respectively. N_{\max} denotes the maximum number of retrieved images. $P(N)$ and $R(N)$ denote the precision and recall values with N images retrieved, respectively.

4.4 Retrieval performance and discussion

The type of features used is very important in image representation. Color, intensity and orientation are the most commonly used visual features in CBIR and object recognition. Retrieval performance and vector dimensionality are the most important factors in CBIR. It is perfectly natural to obtain good retrieval performance by minimizing vector dimensionality. Since the proposed method is a histogram-based method, feature quantization has a strong influence on image retrieval performance. The number of bins of low-level visual features can directly influence the retrieval results.

In this paper, Microsoft Excel 2010 was used for drawing the performance curves, with smooth lines used in Figs. 5, 6, 7 and 8 to make the diagrams reader-friendly.

4.4.1 Evaluation of HSV color space

In order to confirm the contribution of low-level visual features to image retrieval performance, different quantization levels of color, intensity and orientation are used in the gradient-structures histogram in HSV color space and opponent-color space. The opponent-color space is used for gradient-structures detection and further processing, whereas the HSV color space is used for extracting color, intensity and orientation features using the technique of feature quantization.

In the CBIR experiments and comparisons, $\text{bin}(H)$, $\text{bin}(S)$ and $\text{bin}(V)$ denote the numbers of bins for the H , S and V components. In this paper, we adopt the method of gradually increasing the quantity of bins, with $\text{bin}(H) \geq 6$, $\text{bin}(S) \geq 3$ and $\text{bin}(V) \geq 3$ for the feature quantization of the HSV color space; hence, the total number of bins is at least 54 and is gradually increased to 108.

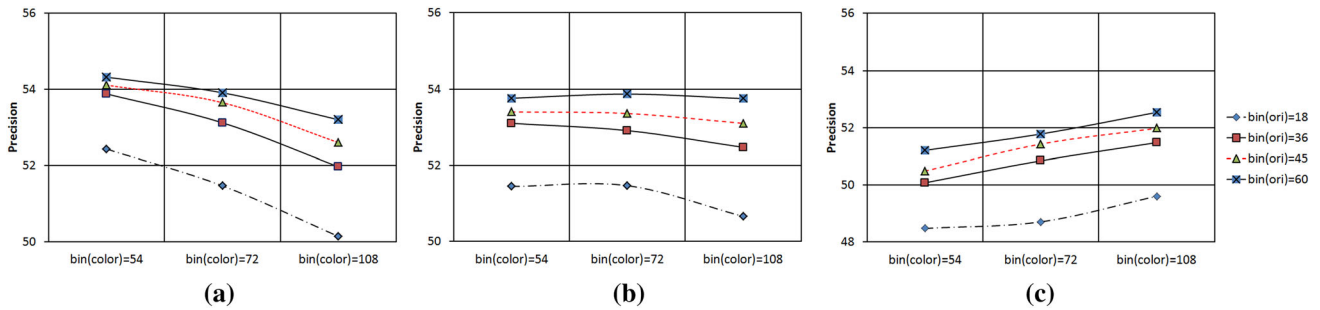


Fig. 5 Average precision using the gradient-structures histogram on the Corel-10K dataset in HSV color space, where bin(gray), bin(ori) and bin(color) denote the quantization levels of intensity, orientation and color, respectively: **a** bin(gray) = 16, **b** bin(gray) = 32, and **c** bin(gray) = 64

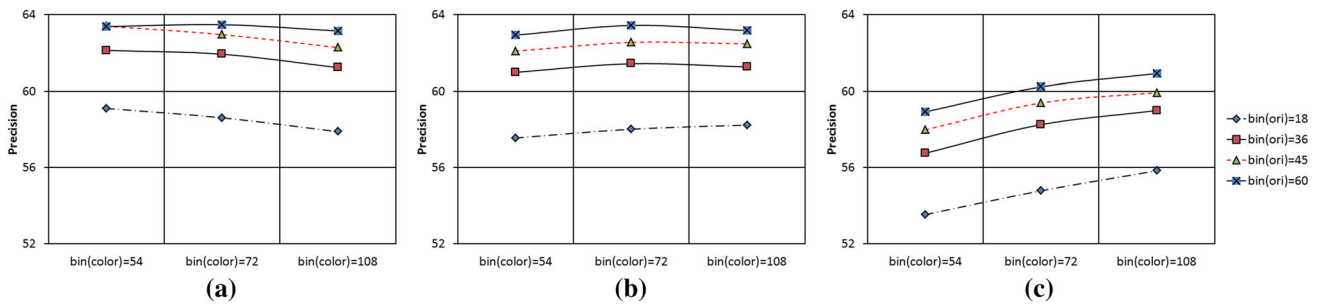


Fig. 6 Average precision using the gradient-structures histogram on the GHIM-10K dataset in HSV color space, where bin(gray), bin(ori) and bin(color) denote the quantization levels of intensity, orientation and color, respectively: **a** bin(gray) = 16, **b** bin(gray) = 32, and **c** bin(gray) = 64

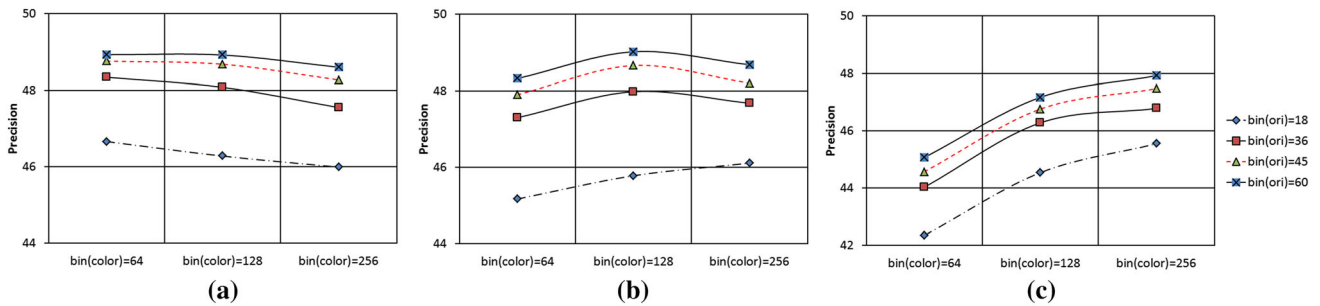


Fig. 7 Average precision using the gradient-structures histogram on the Corel-10K dataset in RGB color space, where bin(gray), bin(ori) and bin(color) denote the quantization levels of intensity, orientation and color, respectively: **a** bin(gray) = 16, **b** bin(gray) = 32, and **c** bin(gray) = 64

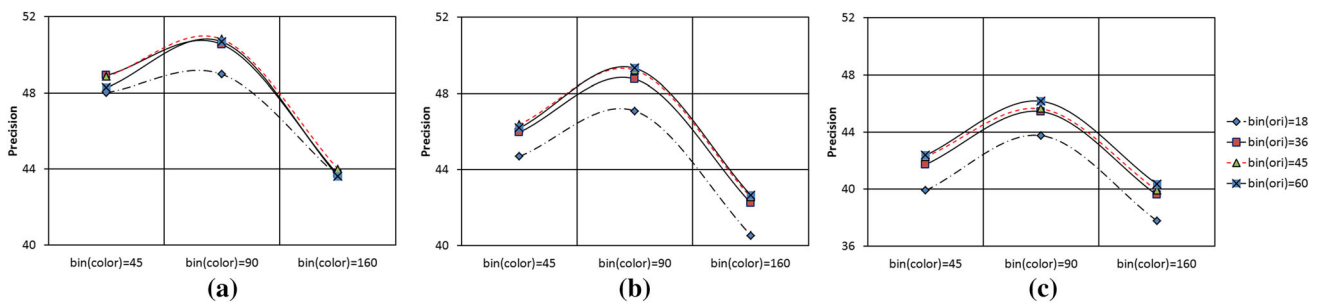


Fig. 8 Average precision using the gradient-structures histogram on Corel-10K dataset in Lab color space, where bin(gray), bin(ori) and bin(color) are denoted as the quantization levels of intensity, orientation and color, respectively: **a** bin(gray) = 16, **b** bin(gray) = 32, and **c** bin(gray) = 64

Figures 5 and 6 illustrate the average retrieval precision of the gradient-structures histogram on both the Corel-10K and GHIM-10K datasets in HSV color space. The best average precision was obtained when $\text{bin}(\text{color}) = 54$, $\text{bin}(\text{ori}) = 60$ and $\text{bin}(\text{gray}) = 16$. The average precision of the GSH was 54.84% on the Corel-10K dataset and 63.38% on the GHIM-10K dataset.

4.4.2 Evaluation of Lab and RGB color spaces

Besides the HSV color space, the RGB and Lab color spaces were also evaluated to determine which is most suitable for use with the proposed algorithm. In Figs. 7 and 8, we can see that the performance of the GSH is better in the Lab color space than in the RGB color space while providing the same or similar retrieval precision.

When using a quantization level of intensity is 16 bins, the precision is about 46–49% in RGB color space and about 43–51% in Lab color space. When the quantization level of intensity is 32 bins, the performance of the GSH is better in Lab color space than in RGB color space. With 64 bins, the performance is better in RGB than in Lab color space.

In the performance comparisons using the Lab and HSV color spaces, it is obvious that the average precision is better using HSV when the quantization level of intensity is fixed at 16 bins. The vector dimensionality of the GSH is higher in Lab color space than in RGB color space, but its precision is lower than in HSV. Considering the vector dimensionality and retrieval precision, we selected the HSV color space for color quantization in the proposed method. The final quantization numbers for color, intensity and orientation were set to 54 bins, 16 bins and 60 bins, respectively.

4.4.3 Evaluation of different distance or similarity metrics

In order to determine performance differences due to the use of different distances or similarity metrics, several of these were adopted in CBIR experiments. As can be seen from Table 1, Canberra gives much better results than

other metrics such as L1, x^2 statistics, Chebyshev, Cosine and histogram intersection. The Chebyshev metric gives the worst results on the two datasets. The computational burden of the Cosine metric is the greatest due to the computation of the dot product.

The GSH uses the merits of low-level visual features by representing the attributes of gradient-structures using a histogram. In the GSH technique, there are many bins with frequencies close to zero. If we apply a histogram intersection and the probability that $\min(\mathbf{T}_i, \mathbf{Q}_i)$ is high, a false match may appear; therefore, a histogram intersection is not suitable for use as a similarity metric for the proposed method. In the Chebyshev metric $D(\mathbf{T}, \mathbf{Q}) := \max(|\mathbf{T}_i - \mathbf{Q}_i|)$, the max operation results in many false matches. In contrast, the Canberra distance is simple to calculate and can be considered as a weighted L1 distance with $1/(1.0 + |\mathbf{T}_i| + |\mathbf{Q}_i|)$ being the weight. Since the same values of $|\mathbf{T}_i - \mathbf{Q}_i|$ can come from different pairs of \mathbf{T}_i and \mathbf{Q}_i , using a weight parameter can reduce these opposing forces.

4.4.4 Performance comparisons

Two image subsets, consisting of 10% of the images from each dataset, were used as query images for the Corel-10K and GHIM-10K datasets. The system performed similarity evaluations on each query image. The performances were evaluated from the average results of all queries using precision and recall metrics. Table 2 shows the comparisons between Bow histogram, CVH, PUD, LBP, CDH and GSH in terms of precision and recall on the two datasets. GSH performs better than the Bow histogram, CVH, LBP and CDH methods. In the CBIR experiments, the vector dimensions of GSH, CDH, CVH, LBP, PUD and Bow histogram were 130, 108, 104, 256, 240 and 1000 bins, respectively. It is clear that the Bow method is much better than the CDH and GSH methods in terms of vector dimensionality.

For a fair comparison with some improved LBP methods and some other methods that work with tree-structured data. All the images in the Corel-1000 dataset were

Table 1 Average retrieval precision and recall performance of the GSH method according to different distance or similarity metrics

Dataset	Performance	Distance or similarity metric					
		Canberra	Cosine	L1	x^2 statistics	Chebyshev	Histogram intersection
Corel-10K	Precision (%)	54.55	33.74	38.87	42.74	29.98	35.73
	Recall (%)	6.55	4.05	4.66	5.13	3.6	4.29
GHIM-10K	Precision (%)	63.11	37.9	43.92	47.96	34.6	41.42
	Recall (%)	1.51	0.91	1.05	1.15	0.83	0.99

Table 2 Comparisons of the BOW histogram, LBP, PUD, CVH, CDH and GSH methods in terms of precision and recall metrics on the Corel-10K and GHIM-10K datasets

Dataset	Performance	Method					
		Bow histogram	LBP	PUD	CVH	CDH	GSH
Corel-10K	Precision (%)	30.36	37.23	50.24	48.58	45.24	54.84
	Recall (%)	3.64	4.47	6.03	5.83	5.43	6.58
GHIM-10K	Precision (%)	39.67	46.89	58.89	57.48	51.79	63.38
	Recall (%)	0.96	1.13	1.41	1.38	1.24	1.52

Table 3 Comparisons with some improved LBP methods in terms of precision/recall and AUC metrics on the Corel-1000 dataset

Method	AUC	Number of retrieved images					
		Precision (%)			Recall (%)		
		10	50	100	10	50	100
GSH	0.4338	82.80	67.10	54.34	8.28	33.55	54.34
Tree2Vector	0.3220	76.67	61.08	50.52	7.67	30.54	50.52
LTP	0.2370	68.01	52.04	42.80	6.80	26.02	42.80
LBP	0.2225	66.47	50.59	41.21	6.65	25.30	41.21
MLSOM	0.2943	69.82	56.90	49.86	6.98	28.45	49.86
LTrPs	0.2378	69.42	52.47	42.44	6.94	26.24	42.24

adopted as query images. Table 3 shows a comparison of the GSH, Tree2Vector, LTP, LBP, MLSOM and LTrPs methods in terms of precision/recall and AUC metrics on the Corel-1000 dataset. In Table 3, all the comparative data for the Tree2Vector, LTP, LBP, MLSOM and LTrPs methods are reported in the conference proceedings of [45]. It is clear that the GSH method greatly outperforms Tree2Vector, LTP, LBP, MLSOM and LTrPs in terms of the metrics used.

In the Bow method, using the SIFT descriptor to extract local features and the implementation of clustering can result in heavy computational and memory costs. It must be pointed out that local-feature detectors are used for reliable object matching from different viewpoints and under different lighting conditions. The images in each category of the Corel-10K and GHIM-10K datasets have similar contents. *Similar contents* mean that the color, texture and shape features are similar such that the images cannot be used for reliable object matching from different viewpoints and under different lighting conditions. This is the main reason why the Bow histogram method does perform well on the Corel-10K and GHIM-10K datasets; however, it can perform excellently in object-based image retrieval and object recognition [25–32].

The CVH method incorporates the advantages of histogram-based methods as it takes into account the spatial information of neighboring colors and edges. Although it performs well, it cannot represent local structures with various widths and orientations, which leads to reduced image retrieval performance in some image classes [5].

The local binary pattern (LBP), which includes the improved LBP methods LTP and LTrPs, is a well-known texture descriptor which can represent the local structures of image or color texture regions. However, it does not have the discriminative power of using edges and bars of certain widths and various orientations.

The CDH was developed for CBIR. Color, orientation and perceptually uniform color differences are encoded as feature representations in a similar manner to that of the human visual system, but the CDH discards intensity information and local structure cues such as gradient structures, which can reduce its ability to describe image content.

The PUD method has a combined color perceptual feature and texton frequency feature [41], and its final corresponding dimensionality is 280 bins. It is derived from the color difference histogram (CDH) [12] and microstructure descriptor (MSD) [13]; however, it is difficult to balance the color perceptual feature and texton frequency feature, which reduces its discriminatory power and increases the vector dimensionality.

Both the Tree2Vector and MLSOM methods work on tree-structured data. Tree2Vector aims to describe the global discriminative information embedded at the same levels of all trees. The limitations of MLSOM include its dependence on designing a specific SOM structure through a careful training process and its incapability to formulate an independent vector for each tree [45], while Tree2Vector can overcome these issues. However, both the Tree2Vector and MLSOM methods do not have the power to discriminate edges and bars of certain widths and various orientations.

Gradient-structures are based on edges and bars of certain widths and various orientations that have opponent colors. Extracting gradient-structures is very useful and beneficial in describing image content, making the discriminatory power of GSH better than that of CDH. Gradient-structures come from consistency in edge orientation. gradient-structures perform a connective function between the orientation selection mechanism and low-level features.

In image representation, using color, edge orientation and intensity as constraints with obvious gradient-structures may improve the retrieval results. It is not necessary to a set weight coefficient for low-level features, but the

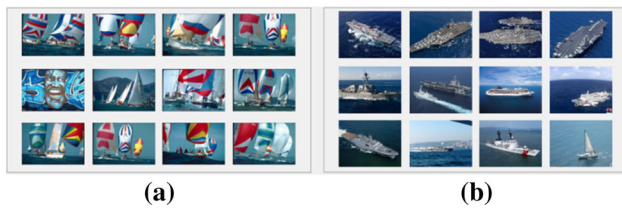


Fig. 9 Two retrieval examples using the GSH method on the Corel-10K and GHIM-10K datasets. The top-left images are the queries and are: **a** a sailboat and **b** a naval ship. The similar images include the query image itself

retrieval results may show a good match with their corresponding low-level features (color, orientation and intensity).

Figures 9a, b show two retrieval examples using the GSH method on the Corel-10K and GHIM-10K datasets. In Fig. 9a, the query is a sailboat image; 11 images were correctly retrieved and ranked within the top 12 images. All the top 12 retrieved images show good matches of texture, color and shape with the query image. In Fig. 9b, the query image is a naval ship, which has obvious shape features and a sea surface as the background. All the returned images were correctly retrieved and ranked within the top 12 images. These included a warship, hospital ship and cruise ship. Two retrieval examples from the sensory effects and a side confirmed that the GSH has the power to discriminate color, texture and shape features.

4.4.5 Limitations of the proposed method

The proposed method has the advantages of being histogram-based and being able to simulate human color perception and the orientation selection mechanism. The gradient-structures perform the function of connecting the orientation selection mechanism with low-level features, which is very useful for describing image contents. However, a limitation of the proposed method is that the gradient-structures are simultaneously sensitive to the represented color, intensity and edge orientation. It is difficult to perfectly balance these three parameters.

Besides, the proposed method does not entirely simulate the visual mechanisms of the human brain. Connecting other visual mechanisms with low-level features according to the principle of visual pathways will be studied in future.

5 Conclusions

Both the orientation selection mechanism and color perception are very important processes of the human brain. In order to extract low-level features by mimicking the orientation selection mechanism and color perception, in this

paper, we propose a detector and discriminative representation system within the CBIR framework. In such a framework, we extract image features by simulating the orientation selection mechanism based on edges and bars of various widths and orientations. In order to mimic color perception well, the HSV color space and opponent-color space were used in the proposed operator and representation.

Experiments were conducted on three datasets and the results compared with those of some existing state-of-the-art methods. The results demonstrate that the GSH method has strong discriminatory power for low-level features (e.g., color, texture and edges) and significantly outperforms the Bow histogram, local binary pattern histogram, perceptual uniform descriptor, color volume histograms, color difference histogram and Tree2Vector methods, as well as some improved LBP methods, in terms of precision and recall and AUC metrics.

In further research, we plan to maintain the existing advantages of our method while introducing deep learning and exploiting other color space properties.

Funding Funding was provided by National Natural Science Foundation of China (Grant No. 61866005) and the project of the Guangxi Natural Science Foundation of China (Grant No. 2018GXNSFAA138017).

Compliance with ethical standards

Conflict of interest We declare that we have no conflict of interest.

References

- Hubel D, Wiesel TN (1962) Receptive fields. Binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106–154
- Marr D, Hildreth E (1980) Theory of edge detection. *Proc R Soc Lond Ser B Biol Sci* 207(1167):187–217
- Manjunath BS, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley, London
- Manjunathi BS, Ma WY (1996) Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell* 18(8):837–842
- Hua Ji-Zhao, Liu Guang-Hai, Song Shu-Xiang (2019) Content-based image retrieval using color volume histograms. *Int J Pattern Recognit Artif Intell* 33(9):1940010
- Singh C, Walia E, Kaur KP (2017) Color texture description with novel local binary patterns for effective image retrieval. *Pattern Recogn* 76:50–68
- Thompson EM, Biasotti S (2018) Description and retrieval of geometric patterns on surface meshes using an edge-based LBP approach. *Pattern Recogn* 82:1–15
- Dubey SR, Singh SK, Singh RK (2016) Multichannel decoded local binary patterns for content-based image retrieval. *IEEE Trans Image Process* 25(9):4018–4032

9. Liu G-H, Yang J-Y (2008) Image retrieval based on the texon co-occurrence matrix. *Pattern Recogn* 41(12):3521–3527
10. Liu G-H, Zhang L et al (2010) Image retrieval based on multi-texon histogram. *Pattern Recogn* 43(7):2380–2389
11. Liu G-H (2016) Content-based image retrieval based on Cauchy density function histogram. In: 12th International conference on natural computation, fuzzy systems and knowledge discovery, pp 506–510
12. Liu G-H, Yang J-Y (2013) Content-based image retrieval using color deference histogram. *Pattern Recogn* 46(1):188–198
13. Liu G-H, Li Z-Y, Zhang L, Xu Y (2011) Image retrieval based on micro-structure descriptor. *Pattern Recognit* 44(9):2123–2133
14. Liu G-H, Yang J-Y, Li ZY (2015) Content-based image retrieval using computational visual attention model. *Pattern Recogn* 48(8):2554–2566
15. Ojala T, Pietikainen M, Maenpaa T (2002) Multi-resolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24(7):971–987
16. Clement M, Kurtz C, Wendling L (2018) Learning spatial relations and shapes for structural object description and scene recognition. *Pattern Recogn* 84:197–210
17. Hong B, Soatto S (2015) Shape matching using multiscale integral invariants. *IEEE Trans Pattern Anal Mach Intell* 37(1):151–160
18. Žunić J, Rosin PL, Ilić V (2018) Disconnectedness: a new moment invariant for multi-component shapes. *Pattern Recogn* 78:91–102
19. Malu G, Elizabeth S, Koshy SM (2018) Circular mesh-based shape and margin descriptor for object detection. *Pattern Recogn* 84:97–111
20. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 60(2):91–110
21. Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. *IEEE Conf Comput Vis Pattern Recognit* 2:506–513
22. Bay H, Tuytelaars T, Gool LV (2006) SURF: speeded up robust features. *Eur Conf Comput Vis* 1:404–417
23. Mikolajczyk K, Tuytelaars T, Schmid C et al (2005) A comparison of affine region detectors. *Int J Comput Vis* 65(1–2):43–72
24. Alahi A, Ortiz R, Vanderghelynst P (2012) FREAK: fast retina keypoint. In: *IEEE conference on computer vision and pattern recognition (CVPR)*
25. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. *IEEE Trans Pattern Anal Mach Intell* 27(10):1615–1630
26. Sivic J, Zisserman A (2009) Efficient visual search of videos cast as text retrieval. *IEEE Trans Pattern Anal Mach Intell* 31(4):591–606
27. van Gemert JC, Veenman CJ, Smeulders AWM, Geusebroek JM (2010) Visual word ambiguity. *IEEE Trans Pattern Anal Mach Intell* 32(7):1271–1283
28. Wang L, Zhou L, Shen C, Liu L, Liu H (2014) A hierarchical word-merging algorithm with class separability measure. *IEEE Trans Pattern Anal Mach Intell* 36(3):417–435
29. Lobel H, Vidal R, Soto A (2015) Learning shared, discriminative, and compact representations for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(11):2218–2231
30. Liu L, Wang L, Shen C (2016) A generalized probabilistic framework for compact codebook creation. *IEEE Trans Pattern Anal Mach Intell* 38(2):224–237
31. Zhou W, Li H, Hong R, Lu Y, Tian Q (2015) BSIFT: toward data-independent codebook for large scale image search. *IEEE Trans Pattern Anal Mach Intell* 24(3):967–979
32. Takahashi T, Kurita T (2015) Mixture of subspaces image representation and compact coding for large-scale image retrieval. *IEEE Trans Pattern Anal Mach Intell* 37(7):1469–1479
33. ImageNet. <http://www.image-net.org>
34. https://en.wikipedia.org/wiki/Color_vision
35. Liu G-H, Yang J-Y (2019) Exploiting color volume and color difference for salient region detection. *IEEE Trans Image Process* 28(1):6–16
36. Burger W, Burge MJ (2009) Principles of digital image processing: core algorithms. Springer, Berlin
37. Gonzalez RC, Woods RE (2018) Digital image processing, 3rd edn. Prentice Hall, Upper Saddle River
38. Treisman A (1980) A feature in integration theory of attention. *Cogn Psychol* 12(1):97–136
39. Lance GN, Williams WT (1967) Mixed-data classificatory programs I—agglomerative systems. *Aust Comput J* 1(1):15–20
40. van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworths, London
41. Liu S, Wu J, Feng L et al (2018) Perceptual uniform descriptor and ranking on manifold for image retrieval. *Inf Sci* 424(2018):235–249
42. Tzelepi M, Tefas A (2018) Deep convolutional learning for content based image retrieval. *Neurocomputing* 275(31):2467–2478
43. Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans Image Process* 9(6):1635–1650
44. Murala S, Maheshwari RP, Balasubramanian R (2012) Local tetra patterns: a new feature descriptor for content-based image retrieval. *IEEE Trans Image Process* 21(5):2874–2886
45. Zhang H, Wang S, Xu X, Chow TWS, Wu QMJ (2018) Tree2Vector: learning a vectoral representation for tree-structured data. *IEEE Trans Neural Netw Learn Syst* 29(11):5304–5318
46. Zhang H, Ji Y, Huang W et al (2018) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. *Neural Comput Appl*. <https://doi.org/10.1007/s00521-018-3579-x>
47. Wang JZ, Li J, Wiederhold G (2001) SIMPLiCity: semantics-sensitive integrated matching for picture libraries. *IEEE Trans Pattern Anal Mach Intell* 23(9):947–963
48. Zhang H, Chow TWS, Wu QMJ (2016) Organizing books and authors by multilayer SOM. *IEEE Trans Neural Netw Learn Syst* 27(12):2537–2550

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.