Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4554911, 12 pages
https://doi.org/10.1155/2022/4554911

Research Article

Sketch-Based Image Retrieval Using Novel Edge Detector and Feature Descriptor

Jianqiang Sheng,[1] **Fei Wang**,[2] **Baoquan Zhao**,[3] **Junkun Jiang**,[3] **Yu Yang**,[4] **and Tie Cai**[1]

[1]*Shenzhen Institute of Information Technology, Shenzhen, China*
[2]*Shantou University, Shantou, China*
[3]*Sun Yat-sen University, Guangzhou, China*
[4]*Shenzhen Securities Information Co., Ltd, Shenzhen, China*

Correspondence should be addressed to Yu Yang; 942773341@qq.com

Received 13 August 2021; Revised 15 November 2021; Accepted 1 December 2021; Published 1 February 2022

Academic Editor: Javier Prieto

With the explosive increase of digital images, intelligent information retrieval systems have become an indispensable tool to facilitate users' information seeking process. Although various kinds of techniques like keyword-/content-based methods have been extensively investigated, how to effectively retrieve relevant images from a large-scale database remains a very challenging task. Recently, with the wide availability of touch screen devices and their associated human-computer interaction technology, sketch-based image retrieval (SBIR) methods have attracted more and more attention. In contrast to keyword-based methods, SBIR allows users to flexibly manifest their information needs into sketches by drawing abstract outlines of an object/scene. Despite its ease and intuitiveness, it is still a nontrivial task to accurately extract and interpret the semantic information from sketches, largely because of the diverse drawing styles of different users. As a consequence, the performance of existing SBIR systems is still far from being satisfactory. In this paper, we introduce a novel sketch image edge feature extraction algorithm to tackle the challenges. Firstly, we propose a Gaussian blur-based multiscale edge extraction (GBME) algorithm to capture more comprehensive and detailed features by continuously superimposing the edge filtering results after Gaussian blur processing. Secondly, we devise a hybrid barycentric feature descriptor (RSB-HOG) that extracts HOG features by randomly sampling points on the edges of a sketch. In addition, we integrate the directional distribution of the barycenters of all sampling points into the feature descriptor and thus improve its representational capability in capturing the semantic information of contours. To examine the efficiency of our method, we carry out extensive experiments on the public Flickr15K dataset. The experimental results indicate that the proposed method is superior to existing peer SBIR systems in terms of retrieval accuracy.

1. Introduction

Over the past decades, digital image as one of the most common media has permeated almost every aspect of our lives. Especially with the current explosion in imaging technologies and the wide availability of affordable imaging devices, the number of images has been soaring at an unprecedented rate on the Internet, and this is posing a very tough challenge: how to retrieve the content of interest from a huge collection of images both efficiently and effectively. To alleviate the suffering, a considerable amount of effort has been devoted to the development of powerful image retrieval systems. Among them, content-based image retrieval (CBIR) approach [1–5] has emerged as an effective solution to address the challenge. In contrast to keyword-based methods, CBIR is capable of achieving better retrieval performance by leveraging features such as colors, texture, shape, and spatial relation. Recently, the rise of touch screen

and its associated human-computer interaction technology make it possible for sketch-based image retrieval (SBIR) [6–9]. Compared to other methodologies, SBIR allows users to retrieve relevant images by drawing a sketch image of their desired object/scene on a touch screen. Such a technique not only provides users with a convenient and intuitive way to formulate a query but narrows the semantic gap between the query and target images and thus is gaining increasing attention in the image retrieval community.

Despite the ease and flexibility of interaction of SBIR, there remain two essential factors that could have a significant impact on its practicality and accuracy of retrieval. (1) Feature representation. This is a process of encoding the key information of a natural image or sketch image into a feature vector, a.k.a. feature descriptor that can be fed into subsequent algorithms to perform specific tasks. In the context of SBIR, an effective feature descriptor can not only work well on natural images but also can be applied to the semantic information extraction of hand-drawn images according to the stroke direction and line continuity of the sketches. In addition, a feature descriptor should also be robust enough to handle the various varieties of image content and be capable of eliminating the ambiguity of sketches caused by the difference in users' painting skills and styles. (2) Feature matching. The feature matching is aimed at measuring the difference between two feature vectors, i.e., the feature similarity between an input sketch query and a nature image in the database. Therefore, a good evaluation metric should be able to accurately and efficiently quantify such a difference.

In this paper, we mainly focus on the image feature representation and introduce a novel feature descriptor for the use of SBIR. The main contributions of this work lie in the following two aspects. Firstly, a Gaussian blur-based multiscale edge extraction (GBME) algorithm is proposed to extract more detailed features by continuously superimposing the edge filtering results after Gaussian blur processing. Secondly, a hybrid barycentric feature descriptor (RSB-HOG) is devised to extract HOG features by randomly sampling points on the edges of a sketch. Since the directional distribution of the barycenters of all sampling points is integrated into the feature descriptor, we thus get a better handle on the edge information of an image. According to our experiments, the proposed method improves the performance of SBIR in terms of retrieval accuracy. More specifically, compared with existing SBIR systems [10–12], it improves the retrieval accuracy by more than 10% on the public Flickr15 dataset. Besides, the proposed feature descriptor can not only accurately capture the semantic information of both sketch and natural images but is also superior to peer methods in dealing with the ambiguity caused by the individual difference in sketch drawing.

The rest of this paper is as follows. In the next section, we introduce the related work of SBIR. Section 3 details the proposed RSB-HOG image feature descriptor and the GBME edge extraction algorithm. To evaluate the effectiveness of our method, we conduct extensive experiments and compare it with peer SBIR systems in Section 4. Section 5 concludes the paper.

2. Related Work

In recent years, researchers have paid a lot of attention to the SBIR and developed various kinds of feature descriptors and matching methods. This section will briefly introduce several related algorithms that are widely used in SBIR systems from the following three aspects: global feature descriptors, local feature descriptors, and feature matching methods.

As a pioneering work, Dalal and Triggs [13] proposed the first SBIR system, query by visual example (QVE). Its concept of feature matching with media databases and the design of the overall system framework has had a profound impact on subsequent research. In order to narrow the semantic gap between hand-drawn sketches and natural images, many studies have been carried out to develop effective feature representation (i.e., the selection of feature descriptors) used in the SBIR system. Existing research on this problem can be roughly divided into two categories: global feature descriptors and local feature descriptors.

For the global feature descriptors, they put particular emphasis on encoding the overall content of an image, such as the color and spatial structure information. Among them, the HOG feature descriptor proposed by Saavedra [14] is one of the representatives. The essence of HOG is to analyze the gradient direction and highlight the texture and edge information of an image. It has been validated as an effective descriptor and widely used in human body detection algorithms. Olivia and Torralba [15] applied the HOG descriptor to the SBIR system and achieved good retrieval results and proposed a novel feature descriptor Soft-Histogram of Edge Local Orientation (S-HELOs) based on HOG. The difference between S-HELO and HOG lies in the fact that HOG samples all pixels of an image, while S-HELO samples the edge area of an image. Furuya and Ohbuchi [12] used the improved HOG algorithm, gradient field HOG (GF-HOG), to retrieve the image based on the stroke color information of the sketch. GF-HOG does not count the HOG features of all pixels in the image but analyzes the HOG features of the pixels that form the edges of the image. Fei et al. [16] proposed the GIST feature descriptors based on the research of Chalechale et al. [6]. GIST is used to describe scene features and transform the image from the spatial domain to the spectral domain. They defined five spectrograms corresponding to different scenes and compared these spectrograms in the feature matching stage. The angular and radial partitioning (ARP) feature descriptor proposed by Gross [17] divides an image by angle and radial and counts the number of valid pixels in the sector after segmentation, while Olivia and Torralba [15] combined HOG and ARP technology and introduced a new feature descriptor, angle, radius, and orientation partition (AROP), which counts the gradient direction of the valid pixels in the fan-shaped area after segmentation. They also developed a new image preprocessing method, with which the result after preprocessing is more similar to the hand-drawn image. These global feature descriptors have the advantages of smaller feature space and faster matching speed than the local feature descriptors. However, due to the missing image details, SBIR systems based on global feature descriptors are not always being satisfactory.

In contrast to the global feature descriptors, the local feature descriptors extract features from the local regions of an image, focusing on the encoding of image details. Bui and Collomosse [10] improved the HOG feature and proposed the local feature descriptor, Skethed feature lines in the Histogram of Oriented Gradient (SHOG). This feature descriptor extracts the HOG feature at the edge of an image to reduce the original HOG feature space, yielding a faster matching speed. Wang et al. [4] proposed an edge feature matching algorithm Edgel index (EI). Compared to other algorithms, it contains more comprehensive information of an image and achieves better matching accuracy. Unfortunately, its feature space is too large, which makes it not suitable for large-scale image datasets. In order to deal with the problem of large-scale image retrieval, Saavedra [14] proposed a local feature descriptor TENSOR based on measuring the edge tensor of the image, which significantly reduces the time overhead of retrieval. Rui et al. [18] proposed the RST-SHELO local feature descriptor, which uses the sketch token (ST) algorithm to extract image edges. The ST algorithm is more reliable than traditional image edge detectors such as the Canny operator. Then, the square root of the S-HELO feature is normalized to improve the retrieval accuracy.

Another research branch in the field of SBIR is aimed at improving the performance of image retrieval by providing effective feature matching solutions. Rui and Collomosse [19] first applied the Bag of Feature (BoF) framework for image retrieval and proved that the framework can reduce the computational complexity of the SBIR system. This system is based on keyword-based retrieval technology, which treats an image as a collection of visual words generated by clustering image feature descriptors, and the number of visual words in each image is regarded as the feature of the image. Bui and Collomosse [10] pointed out that BoF-based SBIR systems eliminate the adverse impact of noise caused by drawing style and the randomness of hand-painting by using the visual words as the final image feature, achieving sound retrieval performance. Yi et al. [20] further explained that this technology reduces the dimension of image feature space and thus improves the storage and retrieval efficiency. Fei et al. [16] proposed a new type of BoF-based image retrieval system using fine-grained sketches. The system is capable of recognizing the detailed semantic clues in a sketch and improves the retrieval accuracy of the SBIR system. Wang et al. [21, 22] classified users into different clusters according to their drawing style and used them as prior information to facilitate image retrieval. There is also a large body of research that seeks to develop deep features for content-based image analysis and retrieval [23–27].

According to the aforementioned analysis, it is not difficult to find that existing studies are more inclined to use local feature descriptors as the image feature representation of hand-drawn sketches. This is because compared to natural images, hand-drawn sketches have little color and spatial structure information, and this makes it difficult to gain the semantic information of a sketch using global feature descriptors. However, using local feature descriptors may introduce a considerable amount of noise and be at the expense of a larger feature space. Therefore, BoF-based schemes are commonly used to cope with these challenges. Inspired by these methodologies, we devise a novel edge extraction algorithm and an RSB-HOG feature descriptor to improve the accuracy of our SBIR system.

3. The Proposed Method

In this section, we introduce the proposed SBIR system from the following three aspects, including Gaussian blur-based multiscale edge extraction (Section 3.2), randomly sampled with barycenter-HOG (Section 3.3), and feature matching (Section 3.4).

3.1. The Framework of SBIR. As shown in Figure 1, the framework of the proposed SBIR system is mainly composed of the following three modules. (1) Edge extraction module. This module is developed to extract edges from natural images using the proposed GBME algorithm, which yields rich edge details by iterating the edge filtering results after blur processing. (2) Deep semantic information extraction module. In this paper, we use the proposed image feature descriptor RSB-HOG to extract the local edge features, which can better overcome the ambiguity caused by the image semantic gap. Then, the BoF is utilized to reduce the dimensionality of the feature space. We obtain the feature vectors of images by establishing the visual vocabularies of the hand-drawn sketch and the image dataset. (3) Feature matching module. This module is designed to measure the similarity between a sketch query and the images in the dataset. We use the top-k most relevant images for performance evaluation. In our implementation, the number of iterations in the GBME algorithm is set to 17, and the size of the visual dictionary in BoF is set to 3000. Such a setting has been proved to be effective in our experiment.

3.2. Gaussian Blur-Based Multiscale Edge Extraction. Due to the intrinsic difference between sketches and natural images, it is difficult to find an effective metric to directly evaluate their similarity. To tackle this issue, a common practice in existing SBIR systems is to formulate an intermediate representation of a natural image using edge detection algorithms and compare the similarity between the obtained edge image and the query sketch. Therefore, the selection of edge detection algorithms could have a great impact on the performance of SBIR systems. Although there are many readily available edge detectors, the efficiency of SBIR systems based on them is still far from being satisfactory due to the aforementioned challenges. In this paper, we propose a novel edge detection method, Gaussian blur-based multiscale edge extraction (GBME), to address this problem.

As shown in Figure 2, (a) is a natural image subset under the theme of "The Tower of London," (b) and (c) are the edge images obtained with the Canny operator and the proposed GBME algorithm, respectively, and (d) shows two hand-drawn sketches that are relevant to the theme while (e) shows two nonrelevant sketches. Obviously, the edge images in Figure 2(c) have more details than those in

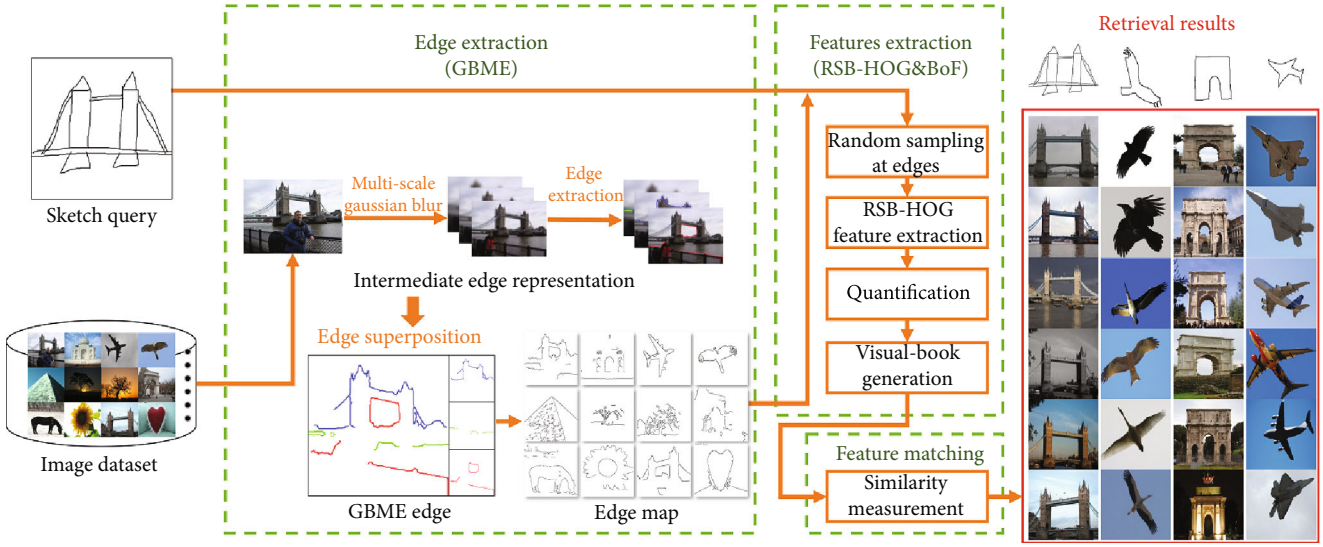

Figure 1: Framework of the proposed SBIR system.

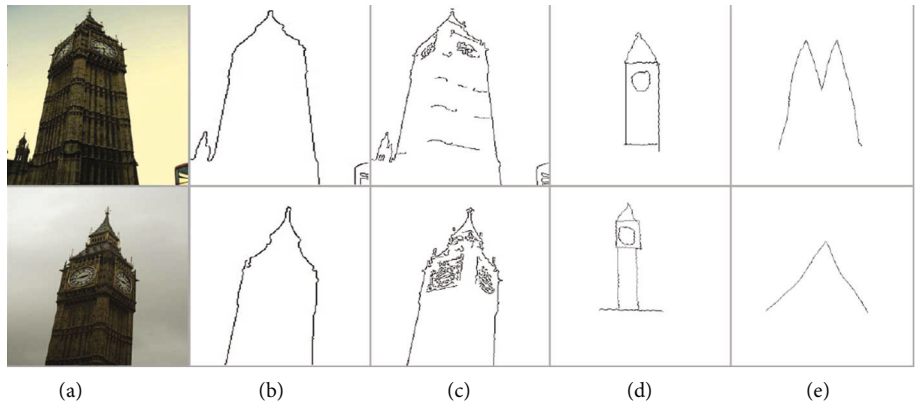

Figure 2: Comparison of two edge extraction algorithms. (a) Two natural images under the theme of "The Tower of London"; (b) the edge images obtained with the Canny operator; (c) the edge images obtained with the proposed GBME algorithm; (d) two hand-drawn sketches that are relevant to the theme; (e) two hand-drawn sketches that are nonrelevant to the theme.

Figure 2(b). Compared to the edge images obtained with our method, the edge images obtained with the Canny operator have the risk of mismatching the nonrelevant sketches in Figure 2(e) due to the absence of the inner structure details of the object, and this may lead to a low retrieval accuracy. The proposed GBME algorithm, however, avoids the occurrence of false matching problems because of its rich edge details. In order to obtain more edge features of the image, the GBME algorithm introduced in this paper adopts an iterative strategy that extracts edges using edge operators from multiscale images after Gaussian blur processing and then takes the sum of the generated edge sets to get an edge image feature with more details.

Unlike existing SBIR systems [10, 12, 28, 29] that directly use edge operators (such as Canny, Laplace, and Sobel) to extract feature edges, the proposed algorithm is intended to simulate the characteristics of the human eyes by using Gaussian blur. A previous study [13] pointed out that the Gaussian bandpass filter is one of the typical models for simulating human vision. The GBME algorithm performs a Gaussian blur operation on a given image with different Gaussian kernel sizes to obtain a collection of images at different scales, which approximates the human eye focusing process from blur to clear after receiving the visual signal, making the extracted edge closer to that of the human perception. As can be observed from Figure 2, the edge images obtained using the proposed GBME algorithm gain more details than the peer method.

3.3. Randomly Sampled with Barycenter-HOG. To extract the semantic information from an image, we also introduce a novel image feature representation, i.e., random sampling point hybrid barycenter image feature descriptor RSB-HOG (randomly sampled with barycenter-HOG), which is an improved algorithm of the global feature descriptor HOG [14]. As shown in Figure 3, the RSB-HOG feature descriptor is randomly sampled at the edges. The sampling points and their center of gravity are visualized in red and blue, respectively. As highlighted in the red bounding box, we use gradient direction of the edge as the HOG feature

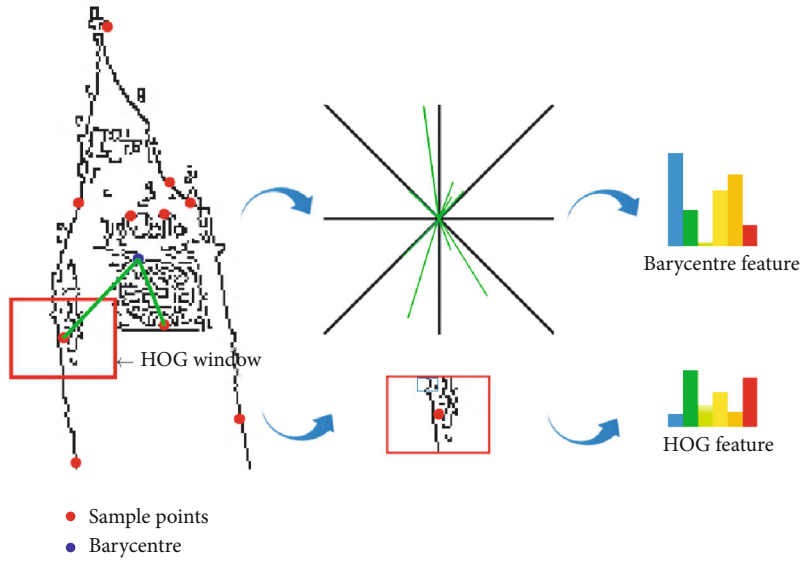

Figure 3: RSB-HOG feature descriptor. The sampling points and their center of gravity are visualized in red and blue, respectively.

of the sampling point (red) and count the directional distribution from the center of gravity (blue dots) of the above sampling points to each sampling point to form a directional distribution histogram as a supplement to the sampling point features. Mathematically, the feature space of RSB-HOG G is given by the following:

$$G = [F, B], \tag{1}$$

$$F = [f_1, f_2, \cdots, f_N], \tag{2}$$

where N is the number of random sampling points, f is the feature vector of a single random sampling point, and B is the feature vector of the center of gravity point.

For the definition of the image edge, since the input of the feature extraction step is the hand-drawn sketch and the image edge set, which is essentially a binary image, therefore, if any pixel in the input image has a grey value, the point can be defined as an edge point. Let $\text{gray}(x, y)$ be the gray value of the pixel at (x, y) in an image, and its edge attribute $\text{IsEdge}(x, y)$ is 1 if it is an edge point, otherwise 0. That is,

$$\text{IsEdge}(x, y) = \begin{cases} 1, & \text{if gray}(x, y) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The steps of RSB-HOG feature extraction are summarised as follows:

(i) Calculate the gray gradient of all pixels in the input image

(ii) Randomly sample the edge point set of the image to obtain the sampling point set $\mathbf{S}$, and the number of points is denoted as β

(iii) Calculate the HOG feature vector f of the sampling point. For the sampling point $s(s \in \mathbf{S})$, the coordinate of the point on the input image is (x, y), and its gray value is $\text{gray}(x, y)$:

(a) Set the window w with this point as the center. The size of the window is set to $w = M \times N$

(b) Set the slider c with a size of $m \times n$ in the window w, and move the slider in the window from top to bottom and from left to right by a given step (α_x, α_y). For each step, record the current position of the slider as $c_i (i \in \{1, 2, \cdots, \lambda\})$, where λ is the total number of times the slider can be moved and is given by the following:

$$\lambda = \left\lceil \frac{M - m}{\alpha_x} \right\rceil \times \left\lceil \frac{N - n}{\alpha_y} \right\rceil. \tag{4}$$

(c) Project the gradient direction of all points in c_i, and use the histogram h_s to count the number of times of the gradient direction grad of each point that hits the calibration direction:

$$h_s(k) = \begin{cases} h_s(k) + 1, & \text{if} d \in \left[\frac{\pi}{6}(k-1), \frac{\pi}{6}k\right], \\ h_s(k), & \text{otherwise,} \end{cases} \tag{5}$$

where $k \in \{1, 2, \cdots, 12\}$.

(d) Normalize the histogram h_s, and the result will be the HOG feature vector f of the sampling point, and its feature dimension is λ

(iv) Calculate the center of gravity of all sampling points: $B = \sum s/\beta$

(v) Project the direction from the center of gravity b to each sampling point s, and use the histogram h_b to count the number of times that each direction d hits the calibration direction:

$$h_b(k) = \begin{cases} h_b(k) + 1, & \text{if } d \in \left[\frac{\pi}{6}(k-1), \frac{\pi}{6}k\right], \\ h_b(k), & \text{otherwise,} \end{cases} \tag{6}$$

where $k \in \{1, 2, \cdots, 12\}$.

(vi) Normalize the histogram h_b to obtain B, and combine F and B to formulate the RSB-HOG feature $G = [F, B]$

In our method, a modest large sampling window size tends to achieve better retrieval performance than smaller ones, since it provides richer details of local information. However, the final accuracy may decrease when the window size exceeds a threshold, i.e., the predefined total number of sampling points. In such cases, some of the candidate points in the window will be discarded randomly, and this may fail to maintain sufficient local information and thus decrease the retrieval performance. In our implementation, the sizes of the sampling window and slider are empirically set to 256×256 and 80×60, respectively, and the number of sampling points of RSB-HOG is set to 500. Since there are 500 sampling points of the whole image, the number of extracted features using HOG is $143 \times 500 = 71{,}500$.

To reduce the image semantic disturbance caused by random sampling and the image feature space, we adopt the BoF method that clusters the local features of each image, and the obtained class centers are defined as visual words in the visual codebook, and each feature in the image will be mapped to the visual dictionary. Therefore, each image can be regarded as an unordered collection of multiple visual words, i.e., the image visual vocabulary, and the number of visual words represents the size of the visual vocabulary. The matching of image features is transformed into the similarity comparison of the image visual vocabulary. Compared with the high-dimensional feature space of the image, the visual vocabulary is a histogram vector whose dimension is determined by the visual dictionary, which consumes less space. The specific steps are as follows:

(1) Use the k-means++ algorithm to cluster the feature sets of all input images, and the generated centroids are visual words. The number of centroids is the total number of visual words of the visual dictionary

(2) Map the feature points of each image to the visual vocabulary obtained in step 1, and count the number of occurrences of the visual vocabulary in each image to construct the visual vocabulary of the image

(3) As did in step 2, construct the visual vocabulary for hand-drawn sketches

(4) Obtain the final visual vocabulary by combining the two visual vocabulary sets using the TF-IDF algorithm [3].

3.4. Feature Matching. In the feature matching module, we take the results obtained in the previous module as the input to evaluate the similarity between a sketch and a natural image. There are many evaluation metrics that can be used to measure the similarity of two feature vectors, including cosine similarity, root mean square distance, average absolute distance, and histogram intersection distance. In our experiment, we found that the cosine similarity metric is superior to other metrics in the context of this work and thus is adopted by us to perform feature matching. The input includes the visual vocabulary of the hand-drawn sketch and the visual vocabulary matrix of the image set, denoted by v_{sketch} and V_{img}, respectively. A row in the visual vocabulary matrix represents the visual vocabulary of an image in the image set, i.e., $v^i_{\text{img}} \in V_{\text{img}}$. The cosine similarity metric COS is defined as follows:

$$\text{COS} = \frac{\sum_{i=1}^{\Phi} v_{\text{sketch}} \times v^i_{\text{img}}}{\sqrt{\left[\sum_{i=1}^{\Phi} (v_{\text{sketch}})^2\right] \times \left[\sum_{i=1}^{\Phi} (v^i_{\text{img}})^2\right]}}, \tag{7}$$

where Φ is the number of natural images in the dataset and λ is the number of visual vocabulary of BoF.

4. Experiment

To verify the efficiency of the proposed method, we conduct extensive experiments and compare it with several peer methods in this section. In addition, we also investigate the influence of different parameters on the accuracy of image retrieval.

4.1. Experiment Setup. We use Flickr15K [20] as the image dataset for comparison, since it is a well-recognized SBIR dataset and widely used by many other existing studies [10, 12, 16, 20, 28]. This dataset is divided into three parts: image set, sketch set, and ground-truth. Among them, the image set contains 14,660 natural images randomly crawled from the Flickr picture sharing website. All images are divided into 60 basic categories such as "pyramid," "bird," and "heart_shape." Figure 4 shows some examples of the natural images in the set. On average, each category contains more than 200 images. The sketch set contains 330 hand-drawn sketches that belong to 33 types of objects drawn by 10 users without professional drawing experience; The ground-truth file contains the information on whether a natural image is related to a sketch. For example, a sketch of a simple circle

Figure 4: Some samples of the images in the Flickr15K dataset.

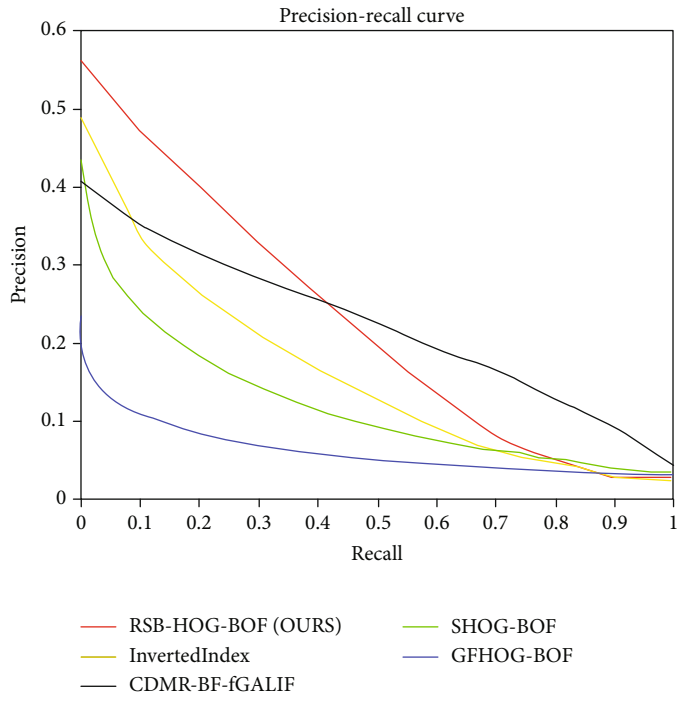

Figure 5: Overall comparison of different SBIR systems.

Table 1: Comparison of MAP in different edge extraction methods.

Method	MAP
SHOG-BOF-EDGE1	19.31%
SHOG-BOF-EDGE2	15.51%
InvertedIndex-EDGE1	20.64%
InvertedIndex-EDGE2	18.21%

Table 2: Comparison of MAP in various methods.

Method	MAP
RSB-HOG-BOF (ours)	25.85%
CDMR-BF-fGALIF	22.50%
InvertedIndex	18.21%
SHOG-BOF	15.51%
GFHOG-BOF	8.27%

should be relevant to the natural images in the categories of "fire balloon," "London eye," "moon," etc.

The experiments were carried out on a PC equipped with an Intel Core i7-4970 processor that operates at a speed of 3.6 GHz and with 16 GB of RAM. The PC runs Windows 10 (64-bit) as its operating system. The development environment used to implement our method is Microsoft Visual Studio 2015. The mean average precision (MAP) and precision-recall (PR curve) are used as the evaluation metrics of the SBIR systems.

4.2. Overall Performance Evaluation of SBIR Systems. To demonstrate the superiority of the proposed method (henceforth referred to as RSB-HOG-BOF), we compare it with peer advanced SBIR systems, including SHOG-BOF [10], GFHOG-BOF [12], InvertedIndex [12], and CDMR-BF-fGALIF [28]. For fair comparison, we obtained the performance on the test data of each system following the same parameter setting that achieves the best performance as reported in the original paper. The performance in terms

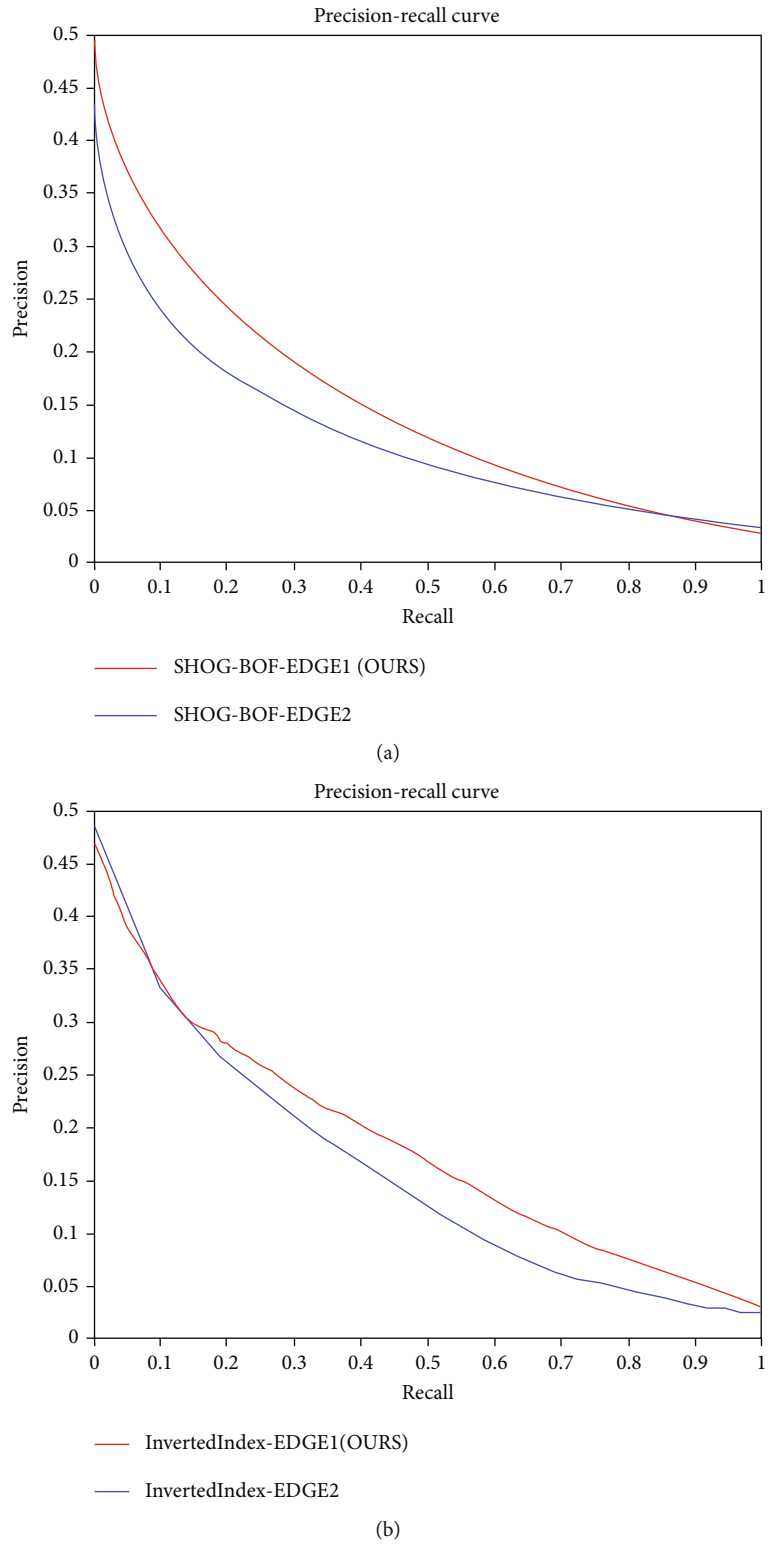

Figure 6: Comparison of PR curves of different edge extraction methods in (a) SHOG-BOF and (b) InvertedIndex SBIR systems.

of MAP and PR of different SBIR systems is illustrated in Figure 5 and Table 1, respectively.

As can be seen from Figure 5, the proposed RSB-HOG-BOF achieves higher image retrieval accuracy than other systems except for CDMR-BF-fGALIF under different recall rates. GFHOG-BOF has a steep drop in accuracy in the recall interval of [0, 0.02], and SHOG-BOF and InvertedIndex also show a significant decrease of accuracy in the recall interval of [0, 0.1]. In contrast, as the recall rate increases, the accuracy of RSB-HOG-BOF decreases smoothly

Table 3: Performance comparison in terms of precision and recall between the Canny edge detector and the proposed GBME in the (a) SHOG-BOF and (b) InvertedIndex SBIR systems.

(a)

Recall@	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
P@SBE2[†]	0.435	0.243	0.184	0.144	0.115	0.094	0.077	0.063	0.051	0.041	0.030
P@SBE1[‡]	0.495	0.313	0.245	0.193	0.153	0.117	0.090	0.070	0.055	0.042	0.030

[†]Precision@SHOG-BOF-EDGE2 (the Canny edge detector). [‡]Precision@SHOG-BOF-EDGE1 (the proposed GBME edge detector).

(b)

Recall@	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
P@IIE2[†]	0.488	0.334	0.262	0.211	0.167	0.126	0.088	0.061	0.047	0.033	0.024
P@IIE1[‡]	0.465	0.339	0.281	0.239	0.203	0.168	0.131	0.102	0.076	0.054	0.030

[†]Precision@InvertedIndex-EDGE2 (the Canny edge detector). [‡]Precision@InvertedIndex-EDGE1 (the proposed GBME edge detector).

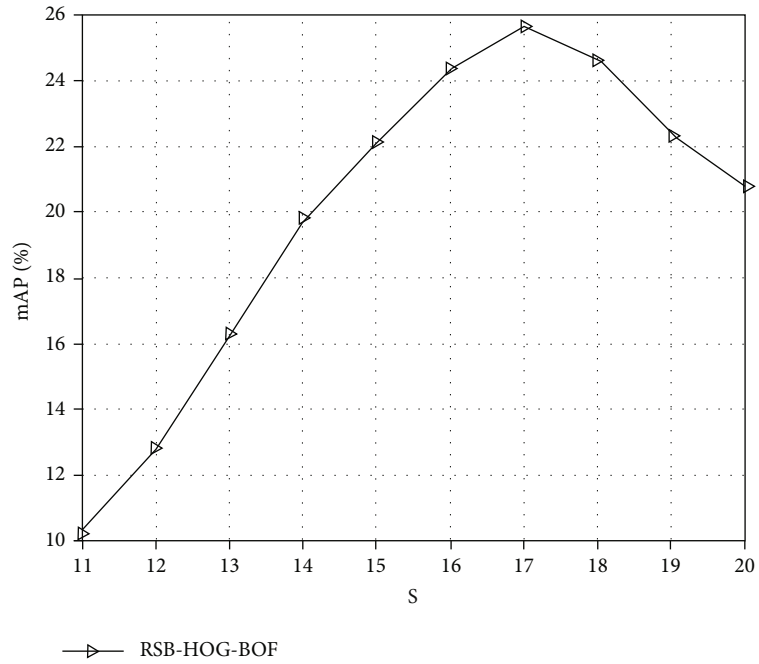

Figure 7: The influence of the number of iterations S on retrieval accuracy.

compared to other systems. In the recall interval [0.4, 1], the accuracy of CDMR-BF-fGALIF surpasses the other four methods.

In addition, the MAP value of each method, as shown in Tables 1 and 2, is calculated as an index to evaluate the discrimination ability of each method. From Table 2, we can see that under the same image set and sketch query, the MAP value of the proposed RSB-HOG-BOF is the highest among all the methods. More specifically, compared to CDMR-BF-fGALIF and InvertedIndex, it improves the performance by 14% and 40.86%, respectively. Besides, GFHOG-BOF underperforms other methods since the GF-HOG is a global feature descriptor and loses too many image details after calculating the tensor at the edge of the image and solving the tensor sparse matrix.

4.3. Efficiency Evaluation of the Proposed GBME Algorithm. To test the effectiveness and versatility of the proposed GBME edge extraction algorithm, we conduct another experiment by comparing it with the Canny edge detector used in two SBIR systems, i.e., the SHOG-BOF system and the InvertedIndex system. We denote the test instances using the GBME edge detector as SHOG-BOF-EDGE1 and InvertedIndex-EDGE1 and the test instances using the Canny edge detector as SHOG-BOF-EDGE2 and InvertedIndex-EDGE2, respectively. Likewise, the MAP and PR curves are used to evaluate the efficiency of different edge detectors. The experimental results are shown in Figure 6 (Table 3) and Table 1. As can be seen from Figure 6(a) and Table 3(a), the performance of SHOG-BOF-EDGE1 is significantly better than that of SHOG-BOF-EDGE2.

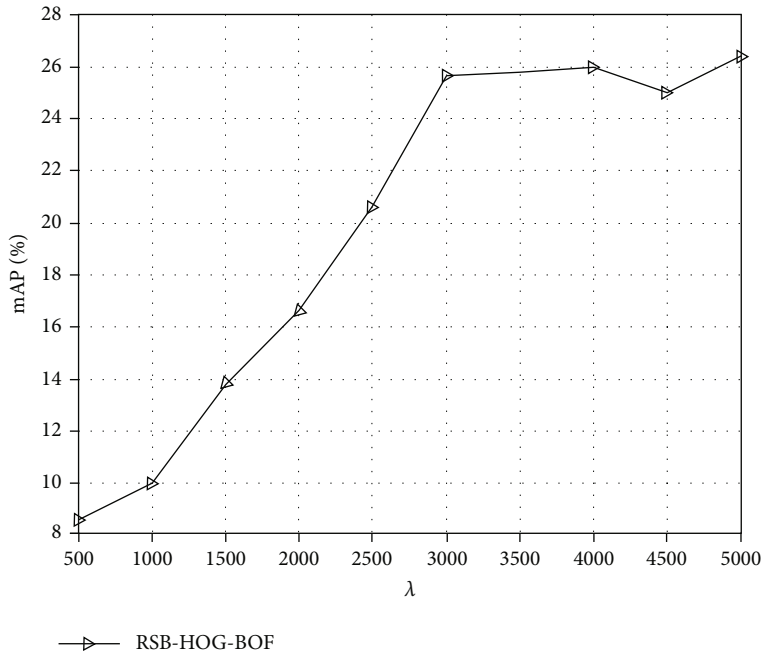

Figure 8: The influence of the number of visual words λ on retrieval accuracy.

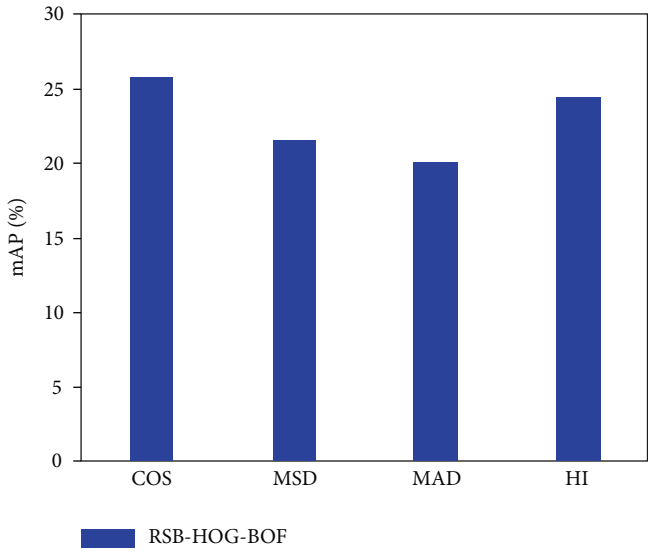

Figure 9: The influence of different similarity metrics on retrieval accuracy.

Additionally, the accuracy of SHOG-BOF-EDGE1 decreases more smoothly in the recall interval of [0, 0.1], while the edge SHOG-BOF-EDGE2 decreases sharply. In Table 1, we can see that SHOG-BOF-EDGE1 improves the MAP by 24.50%, compared to SHOG-BOF-EDGE2. And in Figure 6(b) and Table 3(b), the accuracy of InvertedIndex-EDGE1 and InvertedIndex-EDGE2 is almost equivalent in the recall interval [0, 0.15], and then, the former decreases more slowly than the latter. In terms of MAP value, the former is 20.63% higher than the latter. From the experimental results, we can draw the following conclusions. The proposed image edge extraction algorithm GBME improves the retrieval accuracy of multiple SBIR systems compared to the peer method. It also confirms the assumption that edge images with richer details could benefit the performance of the SBIR systems.

4.4. Impact of Different Parameters on SBIR Performance

4.4.1. The Influence of the Number of Iterations S on Retrieval Accuracy. The proposed GBME algorithm iteratively executes Gaussian operators with different kernel sizes to extract image edges. Therefore, the number of iterations S could have an impact on the performance of SBIR systems. To investigate the impact, we conduct an experiment by constructing an SBIR system with RSB-HOG as its feature descriptor. The system is denoted as RSB-HOG-BOF. We increase the value of S from 11 to 20 with a step size of 1. The MAP is used to measure the retrieval accuracy of the system. The larger the MAP, the better the retrieval performance of the system.

The experimental result is shown in Figure 7. It can be seen from the figure that the parameter S has a great impact on the retrieval accuracy of the system. In the beginning, as the value of S increases, the MAP value of the system gradually increases, which indicates that the increase in the number of iterations results in the detected edges with richer details. As a consequence, the system uses more image information for feature matching, and the retrieval accuracy gradually increases. When the MAP value reaches the maximum, the MAP begins to decrease. Too many iterations will make the detected edges contain more background information of a natural image, and this will have an adverse impact on the retrieval accuracy of the system. Therefore, we set S to 17 in our implementation according to the analysis of this experiment.

4.4.2. The Influence of the Number of Visual Words λ on Retrieval Accuracy. λ is the number of visual words contained in the visual dictionary in the BoF module, which controls the size of the centroid in the clustering operation of the image feature space and the degree of feature reduction in the BoF process. That is, the value of λ is related to the spatial size of the image processed by the module and the computational complexity of the feature similarity calculation in the subsequent steps. The larger size the visual vocabulary, the stronger the ability to express the image, but it also leads to larger feature space and greater computational complexity of the feature matching. Therefore, in order to find a proper value, we use the Flickr15K dataset and the RSB-HOG-BOF system and calculate the MAP under different visual vocabulary settings. In our experiment, we increase the value of λ from 500 to 5000 with a step size of 500. The experimental results are shown in Figure 8. It can be seen from the figure that the MAP value increases rapidly as the value of λ increases at the beginning. When $\lambda > 3000$, the curve gradually becomes smooth. In the interval of [3000, 5000], the increase rate of MAP is less than 1%, and is accompanied by data oscillation. Since too many visual words will increase the complexity of the system in the calculation of similarity, so we choose 3000 as the value of the parameter λ in our implementation.

4.4.3. The Influence of Different Similarity Metrics on Retrieval Accuracy. In the feature matching module of this system, the visual vocabulary of the hand-drawn sketch and the visual vocabulary matrix of the image set are used to calculate the similarity, and the most relevant natural images in the image library are found out for the input sketch query. Therefore, different similarity metrics may harvest different similarities for a given pair of feature vectors, and the final retrieval results will also change. In this experiment, we also use the RSB-HOG-BOF system to examine the difference of different similarity metrics, including cosine similarity (COS), root mean square distance (MSD), average absolute distance (MAD), and histogram intersection distance (HI). The experimental result is shown in Figure 9. As can be seen from the figure, the performance of the system using COS as the similarity metric is the best (25.65%), which is higher than the other methods. We thus use COS as the metric for feature matching.

5. Conclusion

In this paper, we have presented a novel edge extraction algorithm and an effective image feature descriptor to improve the performance of SBIR systems. In the preprocessing stage, the proposed GBME edge detector is first used to convert a natural image into intermediate edge representation. Compared with conventional edge extraction algorithms, it not only captures more essential details but narrows the semantic gap between sketches, edge features, and natural images. In addition, a novel image feature descriptor, RSB-HOG, has also been devised to localize HOG features and add information about the center of gravity of sampling points, and this enhances the ability to distinguish between image feature details and image contours. Furthermore, we use the BoF framework to convert image features into a visual vocabulary to reduce the dimension of feature space and time overhead of feature matching. The experimental results on the public Flickr15K dataset demonstrate the superiority of the proposed method over several existing peer SBIR systems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by the Natural Science Foundation of Guangdong Province (Nos. 2021A1515012302 and 2214050002666), the Scientific Research Start-up Fund of Shantou University (NTF20011), the National Natural Science Foundation of China (61902087), and the General Universities and Colleges Young Innovative Talents Project of Guangdong Province (2019GKQNCX120).

References

[1] N. T. Bani and S. Fekri-Ershad, *Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains*, The Electronic Library, 2019.

[2] J. M. Guo and H. Prasetyo, "Content-based image retrieval using features extracted from halftoning-based block truncation coding," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 3, pp. 1010–1024, 2015.

[3] C. H. Huang, J. Yin, and F. Hou, "A text similarity measurement combining word semantic information with TF-IDF method," *Chinese Journal of Computers*, vol. 34, no. 5, pp. 856–864, 2011.

[4] T. Wang, S. M. Hu, and J. G. Sun, "Image retrieval based on color-spatial feature," *Journal of Software*, vol. 13, no. 10, pp. 2031–2036, 2002.

[5] Xiyu Yang, Xueming Qian, and Yao Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1709–1721, 2015.

[6] A. Chalechale, A. Mertins, and G. Naghdy, "Edge image description using angular radial partitioning," *IEE Proceedings–Vision Image and Signal Processing*, vol. 151, no. 2, pp. 93–101, 2004.

[7] M. Indu and K. V. Kavitha, "Survey on sketch based image retrieval methods," in *International Conference on Circuit*, Nagercoil, India, 2016.

[8] Y. Qian, L. Feng, Y. Z. Song, X. Tao, and C. L. Chen, "Sketch me that shoe," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 799-–7807, Las Vegas, NV, USA, 2016.

[9] C. Yang, C. Wang, L. Zhang, and Z. Lei, "Edgel index for large-scale sketch-based image search," in *CVPR 2011*, Colorado Springs, CO, USA, 2011.

[10] T. Bui and J. Collomosse, "Scalable sketch-based image retrieval using color gradient features," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 1012–1019, Santiago, Chile, 2015.

[11] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: benchmark and bag-of-features descriptors," *IEEE Transactions on Visualization & Computer Graphics*, vol. 17, no. 11, pp. 1624–1636, 2011.

[12] T. Furuya and R. Ohbuchi, "Visual Saliency Weighting and Crossdomain Manifold Ranking for Sketch-Based Image Retrieval," in *MultiMedia Modeling. MMM 2014*, C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, Eds., vol. 8325 of Lecture Notes in Computer Science, Springer, Cham, 2014.

[13] K. Hirata and T. Kato, "Query by visual example," in *International Conference on Extending Database Technology*, pp. 56–71, Berlin, Germany, 1992.

[14] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo)," in *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2015.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[16] W. Fei, S. Lin, H. Wu, R. Wang, and X. Luo, "Data-driven method for sketch-based 3d shape retrieval based on user similar drawstyle recommendation," in *SIGGRAPH ASIA 2016 Posters*, Macao, China, 2016.

[17] M. D. Gross, "Visual languages and visual thinking: sketch based interaction and modeling," in *EUROGRAPHICS Workshop on Sketch-Based Interfaces and Modeling*, New Orleans, Louisiana, USA, 2009.

[18] H. Rui, M. Barnard, and J. P. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *2010 IEEE International Conference on Image Processing*, Hong Kong, China, 2010.

[19] H. Rui and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Computer Vision & Image Understanding*, vol. 117, no. 7, pp. 790–806, 2013.

[20] L. Yi, T. Hospedales, Y. Z. Song, and S. Gong, "Intracategory sketch-based image retrieval by matching deformable part models," in *British Machine Vision Conference, BMVC 2014*, Nottingham, UK, 2014.

[21] Z. Sivic, "Video google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003.

[22] F. Wang, S. Lin, X. Luo, H. Wu, R. Wang, and F. Zhou, "A data-driven approach for sketch-based 3D shape retrieval via similar drawing-style recommendation," *Computer Graphics Forum*, vol. 36, no. 7, pp. 157–166, 2017.

[23] F. Wang, Y. Yang, B. Zhao et al., "Deep 3D shape reconstruction from single-view sketch image," in *2020 8th International Conference on Digital Home (ICDH)*, pp. 184–189, Dalian, China, 2020.

[24] F. Wang, Y. Yang, B. Zhao, D. Jiang, S. Chen, and J. Sheng, "Reconstructing 3D model from single-view sketch with deep neural network," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5577530, 9 pages, 2021.

[25] G. Zhang, J. Yang, Y. Zheng, Y. Wang, Y. Wu, and S. Chen, "Hybrid-attention guided network with multiple resolution features for person re-identification," *Information Sciences*, vol. 578, pp. 525–538, 2021.

[26] B. Zhao, S. Lin, X. Qi, R. Wang, and X. Luo, "A novel approach to automatic detection of presentation slides in educational videos," *Neural Computing and Applications*, vol. 29, no. 5, pp. 1369–1382, 2018.

[27] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," 2021, https://arxiv.org/abs/2105.11722.

[28] B. Li, Y. Lu, A. Godil et al., "A comparison of methods for sketch-based 3D shape retrieval," *Computer Vision and Image Understanding*, vol. 119, pp. 57–80, 2014.

[29] M. Eitz, J. Haysy, and M. Alexa, "How do humans sketch objects?," *ACM Transactions on graphics (TOG)*, vol. 31, no. 4, 2012.