Machine Vision and Applications (2018) 29:1083–1100
https://doi.org/10.1007/s00138-018-0953-8

ORIGINAL PAPER

A survey of sketch-based image retrieval

Yi Li[1] · Wenzhao Li[2]

Received: 18 June 2017 / Revised: 2 May 2018 / Accepted: 18 May 2018 / Published online: 1 September 2018

Abstract
Sketch-based image retrieval (SBIR) has been studied since the early 1990s and has drawn more and more interest recently. Yet, a comprehensive review of the SBIR field is still absent. This survey tries to fill in this gap by reviewing the representative papers studying the SBIR problem. More importantly, this survey tries to answer two important questions which are generally not well discussed: what are the objectives of SBIR, and what is the general methodology of SBIR? The reviewed papers are organized in a chronological way and analyzed by answering these two important questions. As a novel trend, fine-grained SBIR has become the main topic for the recent research. The discussion on it is also integrated. From this survey, we hope that different perspectives can be observed, common values can be discovered and new ideas can be inspired.

Keywords Sketch-based image retrieval · Survey · Abstraction gap · General framework

1 Introduction

Sketching is considered as an intuitive means for expressing thoughts through human history and sketch-like petroglyphs can date back to a much earlier time than texts [16]. In computer vision and computer graphics, sketches also have many applications, such as sketch recognition, sketch synthesis and sketch-based image retrieval. For sketch recognition, it solves the recognition of free-hand sketches [16], professional sketches (e.g., faces) [25,55], symbols [49] and other line-rich objects [5,48] which can be represented by a sketch-like format. Sketch synthesis [37] focuses on synthesizing photos into sketching styles and creating different artistic effects. As for sketch-based image retrieval, it addresses the searching difficulty when texts are not convenient or efficient enough to describe human mind and thus provides an alternative and complementary searching method. This survey places a special concentration on sketch-based image retrieval literature and makes effort to provide a clear overview of the field.

Sketch-based image retrieval (SBIR), which allows the user to search images with a free-hand sketch has been an active research topic under the field of content-based image retrieval (CBIR) for a long time [12,46]. Normally, the input sketch is drawn with a high level of abstraction and just roughly describes the holistic shape and salient local shapes of the searched object/scene. On the other hand, the gallery images are generally realistic photographs or sophisticated art works, which are dramatically different from the input sketch. Above all, the core task of SBIR is to find images which have some object/scene with similar holistic shape and salient local details as the input sketch. An illustration of the SBIR system is shown in Fig. 1.

Substantial papers specifically addressing SBIR have been published since the 1990s. Yet only a few SBIR surveys [1,4,24] have emerged until very recently. Although moderate coverage of the literature and the technical details has been achieved in these works, they share some common drawbacks: (1) only very recent SBIR works have been reviewed, i.e., works after 2005, but there were a considerable amount of works between 1992 and 2005 which made distinct contributions to the field; (2) the major sketch–image differences are not clearly defined—the objectives of SBIR are not well defined; (3) a complete and general SBIR framework that comprises all the necessary modules and addresses the major challenges is absent—the methodology of SBIR is not well defined. Therefore, in this survey, we try to draw a clearer picture of the SBIR field by making some comple-

✉ Yi Li
y.li3@osram.com

Wenzhao Li
liwenzhao@qmul.ac.uk

[1] OSRAM Corporate Technology, Munich, Germany

[2] Queen Mary University of London, London, UK

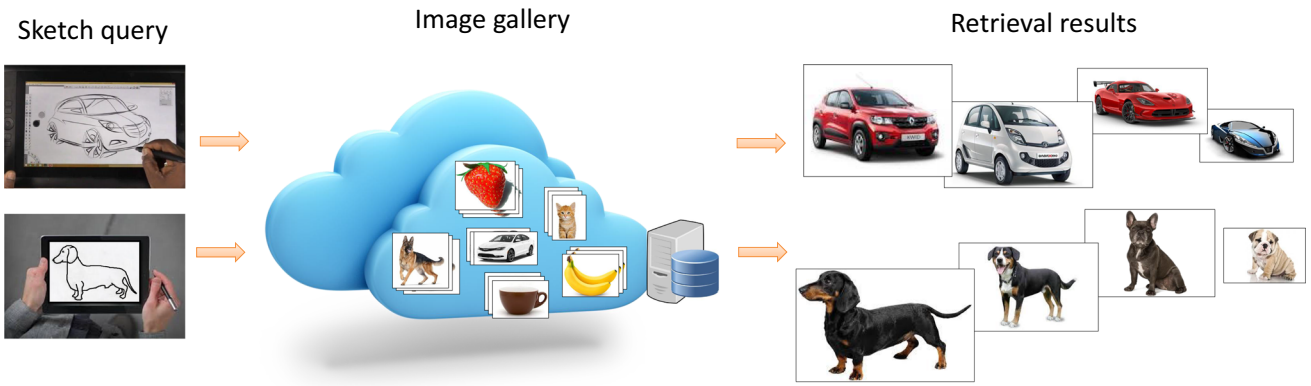

Fig. 1 The concept of the sketch-based image retrieval system

mentary effort to the previous works. More specifically, we review a set of 22 papers from 1992 to 2016 which generally covers the representative SBIR works from the beginning. The sketch–image differences and the corresponding challenges are formally defined and discussed to reflect the clear objectives for SBIR. A general SBIR framework that summarizes all the necessary modules is described, under which the key techniques, strategies and solutions are better organized to offer a comprehensive overview of the methodology. How to use the methodology to achieve the objectives is carefully discussed as well.

The dramatic differences between sketches and images are widely acknowledged yet not well defined in the literature. Various phrases were adopted to coarsely describe them, such as "representational gap" [7], "ambiguity" [28] and "cross-domain" [36,47]. We estimate that finer definitions of these differences would help us understand the problem better and then propose more purposeful solutions. Therefore, we define the sketch–image differences more specifically as below.

Firstly, according to whether the user is sketching the whole scene of the image or an(several) object(s) in the image, we sort the SBIR works into two types: sketching-the-scene type and sketching-the-object type. Then, we define sketch–image differences on three aspects:

1. ***Visual cue imbalance*** sketches only have the holistic shape and salient local shapes (and sometimes symbolic colors), while images have abundant details on shape, color and texture;
2. ***Content imbalance*** sketches normally contain no background, while images can have cluttered background;
3. ***Abstraction gap*** even when a sketch and an edge map are depicting exactly the same object/scene, they still have dramatically different abstraction levels. According to the causes, the *abstraction gap* can be further decomposed into 3 subaspects: random distortion (the randomness in sketch strokes), simplification (missing details) and unrealistic disproportion (caricatured parts being unrealistically bigger or smaller).

The *visual cue imbalance* and the *abstraction gap* exist in both SBIR types, while the *content imbalance* is majorly related to the sketching-the-object type. We argue that these three aspects of differences are the main objectives of SBIR and should be addressed accordingly.

When we review each paper, we mainly discuss about two dimensions:

- the strategies of addressing the sketch–image differences,
- the general SBIR framework.

Under the defined three aspects of sketch–image differences, the different views of the SBIR objectives are revealed and unified. The general SBIR framework is developed incrementally and normally not explicitly discussed in the literature. This survey tries to unfold its developing timeline and summarize all the important modules.

The rest parts of this survey are organized as follows: Sect. 2 reviews all the selected papers in a chronological way, focusing on the strategies of addressing sketch–image differences and the general SBIR framework. Section 3 summarizes the strategies of addressing the sketch–image differences, and Sect. 4 summarizes the general SBIR framework. Section 5 talks about the datasets and evaluation metrics. Finally, Sect. 6 offers some conclusions and some thoughts on the future work.

2 The history of sketch-based image retrieval

The research on SBIR can be traced back to the early 1990s. While the basic concepts are generally consistent across time, the specific research content is highly affected by the corresponding techniques of each period. As a result, we decide

to organize the reviewed works in a chronological way. From this, we hope the developing trends of the SBIR field can be revealed and the promising future directions can be inspired. Based on the shared perspective of the SBIR problem, we divide the time period from 1992 to 2016 into 4 eras and introduce them separately.

On the other hand, as mentioned in the introduction, we analyze all the works through two dimensions: the strategies of addressing sketch–image differences and the general SBIR framework. Along with the strategies of addressing sketch–image differences, we also discuss each work's type and its abilities to address invariances on translation, scale and rotation. There are two types of SBIR in general: the sketching-the-scene type and the sketching-the-object type. Each type is oriented to a specific type of applications. Invariances on translation, scale and rotation affect the user's experience and should be decided according to the applicable scenario. Figuring them out helps us to make clear the designing goals of the SBIR system. For better clarity, we divide the general SBIR framework into two phases: image gallery population phase and sketch retrieval phase. The former phase is the preparation phase and the latter phase is the applying phase.

In the end of the review of each era, tentative summary is provided for the era. A particular focus is imposed onto the strategy to address the *abstraction gap* as we think it is the most important challenge in SBIR. The timeline of the general SBIR framework is also emphasized to highlight the system evolution.

2.1 First era

The first era came in 1992, when Hirata and Kato [26] raised the term of content-based image retrieval for the first time and established a general framework that used sketches to retrieve art paintings. Another version of the same project was published concurrently by Kato et al. [31] which made a clearer manifesto for their framework: "a sketch retrieval method". And they named their whole concept as "QVE (query by visual example)".

To address the *visual cue imbalance*, [26,31] proposed edge extraction module and represented all the gallery images with their edge maps in preprocessing. This module is hence adopted by almost all the following works to tackle the *visual cue imbalance*. To compare the similarity between a sketch and an image edge map, they firstly divided the normalized sketch and edge map into the same number of blocks and then compared the block to block correlations which were computed on raw pixel values directly without extracting any feature. We categorize this kind of metric which compares raw pixel values directly as pixel-based metric. The similarity comparison based on the similarity metrics is then another module. The blocks were shifted within a small range during the comparison, so the similarity metric tolerates translation and random distortion to some extent. After using the similarity metric to compare the query sketch with all the gallery images, they reordered and displayed the gallery images according to their similarities to the sketch in descending order, and we define this procedure as the gallery displaying module. As for the SBIR type, QVE had chosen the sketching-the-scene type.

Right after this thread of works, the QBIC (Query By Image Content) project [21,42] was published by IBM research. As another early and important work in content-based image retrieval, the QBIC project comprised SBIR, yet also included image retrieval from color and texture indications and silhouette shape retrieval. Free-hand sketches/shapes, various colors and synthesized textures could be drawn/selected from an interface. The QBIC project's sketch retrieval pipeline was directly derived from the QVE works [26,31] with no modifications.

Summary The first era lasted from 1992 to 1994 and the symbols are the establishment of some basic modules of the general SBIR framework, the employment of pixel-based metric and the unity on sketching-the-scene type. The only strategy in the first era to address the *abstraction gap* is the block-shifting scheme. It can tolerate the random distortion but cannot address the other two subaspects.

By the end of the first era, the early works had drawn a preliminary picture for the general SBIR framework which can be summarized as follows:

- Population phase: (1) edge extraction;
- Retrieval phase: (1) similarity comparison, (2) gallery displaying.

2.2 Second era

From the second era, researchers started to put more thoughts into the SBIR objectives, i.e., the sketch–image differences. Del Bimbo et al. published several relevant papers [13–15] to describe an elastic matching approach, which deforms the query sketch according to the image object and uses both the deformation energy and the matching extent between the deformed sketch and the image (the overlapping between the sketch pixels and the image edge pixels) to perform similarity comparison. The elastic matching was expected to approximate the human visual perception, and during the deformation process both random distortion and unrealistic disproportion could be addressed. Del Bimbo et al. [14] was the first work in the literature to study the sketching-the-object type, and they added another module to the general SBIR framework which is object localization (yet only necessary for the sketching-the-object type). They [13–15] manually selected some rectangular interesting areas in the

images which were the bounding boxes of the objects. Multi-object query was considered by using a signature file for each sketch and each image to encode the rough spatial relationships between the objects. Corresponding interesting areas of the sketch and the image would be compared only if the signature files of the sketch and the image had matched. Due to the use of rectangular interesting areas, their method is invariant to translation and handles the *content imbalance* (only interesting objects were selected and compared). They also achieved invariance on scale by normalizing the sketch and image objects' rectangles to the same size. It is worth noting that Del Bimbo and Pala [13] for the first time quantitatively evaluated sketch's ability to rank similar images using ranked image list constructed from human rated sketch–image pairs.

So far, all the reviewed works have employed pixel-based similarity metrics. But a pixel-based metric usually is computationally costly and still too rigid to address random distortions. Being aware of this, the feature extraction module was consequently raised by the follow-up works to extract various types of features which were robust to edge variations and more efficient to compare with. Chans et al. [9] believed that the users tended to ignore details when drawing the sketches, and therefore proposed a curvelet model to extract and encode the prominent edge segments of the images. The curvelet representation was intrinsically robust to random distortion and simplification. It was also designed to be invariant to translation within a region. Rajendran and Chang [45] employed a multi-scale representation for edge maps to address the variations of the level of detail of human sketches. Namely, coarser scales of the representation were used to encode only long and prominent edges corresponding to the holistic structure, but finer scales were used to encode also short and weak edges depicting the details. They believed that some combination of the scales could contain the similar amount of details as the query sketch. Invariances on translation, scale and rotation were achieved by using a curvature-direction representation. Nevertheless, their algorithm is just suitable for the sketches of simple shapes and the images with dominant object(s) and clean background. Chalechale et al. [8] proposed an angular partitioning of abstract image (APAI) representation to bridge the *abstraction gap*. Firstly, the abstract images were obtained which would keep the strong edges in the image and derive a thinned version from the sketch, thus addressing the simplification and sketch denoising. The angular partitioning feature was then extracted by covering the abstract image with a surrounding circle and separating the circle into even angular partitions (slices). The number of the edge points in each partition was used to represent that partition. This feature is actually a naive version of the shape context [2] without radial bins. This angular partitioning feature is scale invariant and is further made rotation invariant by calculating its 1-D discrete Fourier transform. However, the rotation invariance is only valid when the rotation angle is an integral multiple of the partition angle. The feature is also robust against random distortion since the distortion would not dramatically change the edge points number in the partition given a decent partition angle. Its translation invariance is achieved by obtaining the bounding boxes of the sketch content and the image content and just comparing the contents in the bounding boxes.

Summary The second era was approximately between 1994 and 2005, and the symbols are the considerations of both the *abstraction gap* and the invariances on several properties. Also in this era, the researchers started to study the SBIR of the sketching-the-object type. Contrasting to the first era, all the three subaspects of the *abstraction gap* were addressed in the second era.

- For the random distortion, the elastic matching [13–15] tried to reverse the distortion with an edge deformation method; the curvelet model [9] reduced the random distortion in the process of extracting the salient edge segments; by just counting the pixel number in each bin, the APAI [8] had achieved invariance on random distortion inside each bin, but also had lost quite a lot of local structures.
- For the simplification, both Chans et al. [9] and Chalechale et al. [8] kept the salient edges as the simplified representation for the images, yet Chans et al. [9] focused more on this process and proposed several steps to regulate the extracted edges while Chalechale et al. [8] just used the thresholds to control the amount of the edges. Rajendran and Chang [45] particularly employed a multi-scale representation to simulate different levels of human simplification.
- For the unrealistic disproportion, only the elastic matching [13–15] could address the situation when the sketch is a contour, but it may not be effective when the sketch has more complicated inner structures.

In the end of the second era, the general SBIR framework appeared as:

- Population phase: (1) object localization (for sketching-the-object type), (2) edge extraction, (3) feature extraction;
- Retrieval phase: (1) feature extraction, (2) similarity comparison, (3) gallery displaying.

2.3 Third era

In the third era, influenced by the explosion of the Internet data volume, the community favored to employ large-scale datasets and investigate features and schemes suitable for processing these datasets efficiently. Eitz et al. [17,18] evalu-

ated histograms of oriented gradients (HOG) [11] and Tensor descriptor [33] against angular radial partitioning (ARP) [8] (yet we think this is a misinterpreted version from Eitz et al., since as introduced earlier, the APAI proposed by Chalechale et al. [8] does not have radial bins) and edge histogram descriptor (EHD) [40], all in a global feature fashion (the feature that represents the whole image). The results concluded that as global features, HOG and Tensor significantly outperformed ARP and EHD, with Tensor being slightly better than HOG. Another two interesting comparisons were also conducted in this work. Firstly, a binary query-specific mask was employed to get rid of the image features whose locations corresponded to the blank area of the query sketch. The retrieval performance using descriptors with masks significantly outperformed the performance without masks. And this is the direct support to the importance of addressing the *content imbalance*. Secondly, they compared traced sketches (traced on top of the image) with memory sketches (sketched from memory without the image), and proved that the traced sketches obtained better retrieval results. This conclusion is expectable, but we would like to argue here that studying the retrieval with memory sketches is more desirable since the users of a search engine normally have no images to trace on. To cope with the large-scale datasets, another module was added to the general SBIR framework in this era: gallery indexing. A k-means tree [23] and best-bin-first strategy [41] were employed for indexing the descriptors. The k-means tree recursively subdivides the descriptors into k clusters with k-means clustering and thus forms a hierarchy of clusters. All the images in one leaf cluster share similar appearances, which could facilitate coherent image retrieval. The best-bin-first strategy is an approximate approach which greedily finds the nearest cluster. Although not generating global optimal results, the combination of k-means tree and best-bin-first strategy could accelerate the retrieval by several orders of magnitude. Due to the global feature fashion, this work cannot cope with translation, scale and rotation (except the ARP which is invariant to rotations of certain degrees). Yet with the grid division scheme in feature computation, random distortion could be tolerated in each grid cell, since the overall statistics of the visual cues in each cell would not change too much given that a moderate level of random distortion is present. Right after, by including both the global features and several popular local features, Eitz et al. [19] proposed a benchmark for SBIR. The local features included shape context (SC) [2], spark feature [19] and HOG. Bag-of-words (BOW) representation [51] was used to encode the features. For indexing, the standard inverted index that indexes images by the visual words [60] was directly employed. In their comparison, although the local feature SHOG (HOG computed on the slightly blurred Canny [6] edge map) obtained the best performance, the other local features were outperformed by the global features. It is noteworthy that, in this work, the sketch's power to rank similar images was quantitatively evaluated again after Del Bimbo and Pala [13]. Ranked image list constructed from manually rated sketch–image pairs was used for retrieval evaluation. The involvement of the BOW representation brings the invariance on translation. The grid division scheme of the features could offer invariance on random distortion as discussed before, and the quantization procedure of the BOW also could offer invariance on random distortion since similarly deformed feature variants are quantized into the same word. Simplification is considered in this work through controlling the thresholds for generating the edge map and employing slightly blurred version of the edge map.

In the meantime, Hu et al. [27] was doing a similar evaluation for local features and the BOW representation, yet with a slightly different set of features. Specially, they proposed to compute a gradient field image for the edge map in advance of computing the HOG feature, and named this feature as gradient field HOG (GF-HOG). The gradient field is composed of interpolations of edge pixels, so essentially, it expands the influences of the prominent edges and increases the generality of the representation. In a later journal version [28], Hu and Collomosse increased their feature set which finally included GF-HOG, HOG, scale-invariant feature transform (SIFT) [38,39], self-similarity descriptor (SSIM) [50], SC and Tensor descriptor (global feature). Their results confirmed that local features, especially GF-HOG and HOG, were obviously better than global features in SBIR. However, their evaluation was category level retrieval, which did not take care of the visual similarity between the sketch and the image, so it provided relatively limited insights on the visual similarity discrimination power of different features. The k-d tree [3] was mentioned in their work for indexing, which organizes the k-dimensional points by partitioning the space and is an important data structure for nearest neighbor search. Nonetheless, the k-d tree scheme can only work with Minkowski distance [3], which is not ideal when other distance metrics are desired. This work could also cope with invariances on translation and random distortion as the features employed and the BOW representation. The gradient field has the effect of simplifying the edge maps and make the edge maps closer to the sketches. Continuing the evaluation of the local features, Hu et al. [29] also considered the *content imbalance*. Since the desired object could be contained in a cluttered scene, they employed a hierarchical segmentation algorithm to decompose the image into coarse-to-fine regions in a recursive way. And they assumed that one of the regions might contain the just right region of the object without too much cluttered background involved. BOW representations were computed for the sketch query and all the regions of each image, and in each image the region with the smallest distance to the query sketch was chosen to represent that image in the retrieval. Significant performance improvement

was reported comparing to the approaches without region segmentation.

Differently, another pixel-based metric was proposed in this era by Cao et al. [7]. They proposed a similarity comparison method based on oriented Chamfer matching (OCM) [7]. The OCM seeks the nearest edge pixel (edgel) in the sketch for each edgel in the image, requiring the quantified orientations of both pixels being the same. The distances between these kinds of edgel pairs were accumulated and normalized to be the final distance. Their indexing scheme was like the inverted index scheme but used edgels instead of feature visual words as entries to index images. All the sketches and images were normalized to the same size, and each edgel entry (with a specific position and a specific edge direction) indexed a list of images which contained that edgel. A Hit map was generated for each sketch as the final representation used for retrieval, and it essentially expands the width of the strokes. By this means, the random distortion is tolerated. However, due to the rigidity of Chamfer matching (which is discussed in [56]) and the normalization needed for the indexing, the system could only retrieve objects at almost the exact position as the sketch and only small extent of local random distortions could be allowed. A very similar work to Chans et al. [9] was proposed by Parui and Mittal [44] that also encoded prominent edge segments in the images. In their work, the salient contours were extracted and decomposed into straight line-like segments which were again formed into a set of long chains. Each chain was encoded by the length ratios and the joint angles of the adjacent segments, and this representation is invariant to translation, scale and rotation. Finally, a fast dynamic programming-based approximate substring matching algorithm was used to match two chains. The similarity between a sketch and an image was accumulated from the scores of the matched chain pairs. And a geometric consistency check for the matched chain pairs was further performed to enforce the holistic structure similarity between the sketch and the image. A hierarchical k-medoids tree was used to index the images. This structure highly resembles the k-means tree, and was adopted to cope with the variable lengths of the chains through k-medoids clustering [43]. In each leaf cluster, the images have at least one chain close to the medoid chain of that cluster. The extracted salient contours are estimated to have a high chance of belonging to an object boundary, and thus could address the *content imbalance* and the simplification. The process of forming the chains has the regularization effect and could tackle the random distortion. However, the straight line-like segments used to compose the chains and the simple representation employed for the chains have lost a considerable amount of local shape information, so the visual similarity between the query sketch and the retrieved images may not be ideal.

Summary The third era started from 2009 and lasted to 2014. Its most significant symbols are the evaluation of various features and the effort of working toward large-scale datasets. Gallery indexing was proposed for the first time in this era and many different indexing schemes were explored by different works. In addition, two subaspects of the *abstraction gap* were addressed in the third era.

- For the random distortion, several strategies existed in the third era. The grid division scheme of the features [17–19,27–29] could tolerate moderate random distortion in each cell, while the BOW representation [19,27–29] could further merge the similarly distorted stroke variants. The gradient field [27,28] and the Hit map [7] both have big potential to address the random distortion since the widths of the sketch strokes are essentially expanded to tolerate the distortions. Finally, the regulated chains (salient edge segments) [44] were used again like in the second era for the random distortion.
- For the simplification, two strategies were inherited from the second era. The first one which was also the most common one [19,27–29] still used the thresholds of the edge detection to control the quantity of the edges. And the second one [44] achieved simplification through the regulated salient edge segments extraction process.

The third era had completed the general SBIR framework as:

- Population phase: (1) object localization (for sketching-the-object type), (2) edge extraction, (3) feature extraction, (4) gallery indexing;
- Retrieval phase: (1) feature extraction, (2) similarity comparison, (3) gallery displaying.

2.4 Fourth era

The emerging fourth era has begun since 2014. In the SBIR literature introduced so far, the sketches' superior power to distinguish intra-category shape variations as opposed to texts was generally ignored except in [13] (Eitz et al. [19] also considered this yet did not focus on the intra-category scenario). However, in [13], Del Bimbo and Pala just experimented with 3 sketch bottle variations without explicitly discussing this particular ability of the sketches. Li et al. [36] first noticed this and raised the concept of "fine-grained" SBIR to highlight the value of using sketches to distinguish intra-category object variations. And the concentration on fine-grained SBIR and the sketching-the-object type have hereafter become the symbols of the fourth era.

To perform reliable object localization and address the *abstraction gap*, the deformable part-based model (DPM) [22] was adopted by Li et al. The DPM served as both an object detector and a representation to bridge the *abstraction*

gap. Each DPM had a two-layer structure: the root and the parts. While its root layer was exploited to encode the holistic pose and address the simplification (as it was computed in lower resolution), its part layer was utilized to encode the configuration and the appearances of the parts. The HOG feature employed in DPM offered their method the ability to cope with local random distortions. The detection function of DPM made their method invariant on translation and scale, and capable to address *content imbalance*. However, since they were using DPMs without part supervision, the obtained parts varied a bit in different domains (sketch parts were not totally the same as the image parts). Graph matching was employed to solve the part correspondences between DPMs from two domains, and a similarity metric that considered both the root and the parts was used for similarity comparison. In the experiments, Li et al. defined the fine-grained similarity between a sketch and an image as 4 aspects: viewpoint, zoom, configuration and body feature. A cross-domain dataset that was merged from the TU-Berlin dataset and the very challenging PASCAL VOC [20] image dataset was used for evaluation. Sketch–image pairs from a small portion of that dataset were manually annotated for fine-grained similarity based on the 4 aspects. Similar to Del Bimbo and Pala [13] and Eitz et al. [19], they also quantitatively evaluated sketches' ability to rank similar images, yet with a scoring scheme.

Two more works [47,58] have adopted the fine-grained SBIR concept, and both of them have used deep learning techniques (specifically triplet network [57]). A very strict fine-grained similarity definition that assumes for each sketch there is only one correct image instance has been utilized by both of them, and it has been named as fine-grained instance-level retrieval. Yu et al. [58] have proposed a dataset containing 716 sketch–image pairs from 2 categories (shoes and chairs) and 32,000 annotated triplets. In their triplet network architecture, all the 3 branches have shared the same set of weights and each branch has been initialized into a state-of-the-art sketch recognition network improved from Sketch-a-Net [59]. Consecutively, a pre-training on automatic generated triplets and a fine-tuning on human annotated triplets have been conducted to train the final triplet model. They have compared their model to Siamese network with contrastive loss [10] and RankSVM [30] with multiple features and have proved their proposed triplet architecture worked best. In the convolutional neural network (CNN) [34], the convolutional layers and max-pooling layers together can cope with random distortion and translation, and the max-pooling layers can also address the simplification by offering abstraction for the objects in the pooling. Sangkloy et al. [47] have done a very similar work concurrently, which yet still has had some differences with Yu et al. [58]. Firstly, they have proposed a much larger dataset that contained 75,471 sketches and 12,500 images from 125 categories (with 100 images for each category and more than 5 sketches corresponding to each image). Secondly, the classification loss has been combined with triplet ranking loss in their optimum setting. Thirdly, while Yu et al. have been employing a customized version of AlexNet architecture [34] called Sketch-a-Net [59] specifically tailed for sketches, Sangkloy et al.'s best model has been directly employing a standard GoogLeNet architecture [54] which is much deeper than AlexNet. Fourthly, Sangkloy et al. have generated each triplet with a sketch, a matching image and a non-matching image from ground truth, but Yu et al. have used both triplets automatically generated according to feature distances and triplets manually annotated by human subjects. Therefore, Sangkloy et al.'s triplets overall are more faithful while Yu et al.'s triplets contains more fine-grained supervision. Exceptionally in the literature, Sangkloy et al. has not employed any explicit edge detector, as they have argued that the detected edges might not represent the human sketching results well and the deep networks could learn such a transformation from data. Although this is an interesting hypothesis, more supporting experiments are needed to verify it.

Very recently, Song et al. [53] have introduced attributes into SBIR, which is an alternative path to cross the *abstraction gap* from a semantic perspective. By integrating the CNN model, they have proposed a multitask ranking network that updates the triplet network constructed in [58]. Apart from the triplet ranking main task, two new tasks—attribute prediction and attribute ranking—have been added to accomplish the final triplet ranking. The three tasks share the same triplet network and contribute together to the overall loss function. An automatic triplet generation strategy that utilizes attribute similarity and ImageNet CNN features has also been proposed to ease the burden of massive triplets generation.

Summary The most profound symbols of the fourth era are the fine-grained SBIR and the exploitation of deep learning. The fine-grained SBIR emphasizes on distinguishing intra-category variations, which is a very valuable scenario especially in commercial products search. The fourth era again has mainly tackled two subaspects of the *abstraction gap*, but also has offered a semantic approach to cross the gap which is using attributes.

- For the random distortion, Li et al. [36] still relied on the grid division scheme of the HOG feature. But deep learning-based works [47,58] have addressed the random distortion in the convolution and the max-pooling, as abstraction is involved in these processes.
- For the simplification, DPM offers a layer at lower resolution (root) to reflect abstraction. And the max-pooling layers abstract the object multiple times in the feature extraction process.

Table 1 The analysis of the strategies of addressing sketch–image differences, the invariances on translation, scale and rotation, and the type of SBIR for each reviewed work

Year	Author	Visual cue imbalance	Content imbalance	Abstraction gap			Trans.	Scale	Rot.	Type
				Dist.	Simp.	Disp.				
1992	Hirata and Kato [26] ([31])	✓		✓			✓			S
1994	Faloutsos et al. [21] ([42])	✓		✓			✓			S
1997	Del Bimbo and Pala [13] ([14,15])	✓	✓	✓		✓	✓	✓		O
1997	Chans et al. [9]	✓		✓	✓		✓			S
2000	Rajendran and Chang [45]	✓			✓		✓	✓	✓	S
2005	Chalechale et al. [8]	✓		✓	✓		✓	✓	✓	S
2010	Eitz et al. [18] ([17])	✓	✓	✓						O,S
2011	Eitz et al. [19]	✓		✓	✓		✓			S
2011	Hu et al. [29]	✓	✓	✓	✓		✓			O
2011	Cao et al. [7]	✓		✓						S
2013	Hu and Collomosse [28] ([27])	✓		✓	✓		✓			S
2014	Parui and Mittal [44]	✓	✓	✓	✓		✓	✓	✓	O
2014	Li et al. [36]	✓	✓	✓	✓		✓	✓		O
2016	Yu et al. [58]	✓		✓	✓		✓			O
2016	Sangkloy et al. [47]			✓	✓		✓			O
2016	Song et al. [53]	✓		✓	✓		✓			O
No.	16	15	5	15	11	1	14	5	3	S(9),O(8)

If the reviewed work is able to address one aspect, the corresponding cell is checked. The eras are separated by double horizontal lines. The abbreviations used in the table are: translation (Trans.), rotation (Rot.), random distortion (Dist.), simplification (Simp.), unrealistic disproportion (Disp.), sketching-the-scene (S) and sketching-the-object (O). The works stated in parentheses are the precedent or succedent works that have considerable content overlap

3 Summary on the strategies of addressing the sketch–image differences

After introducing the history of SBIR, we revisit one of our main focuses. Addressing sketch–image differences is the main objective for SBIR. In this section, we go through each aspect of the sketch–image differences and summarize to what extent each aspect is addressed throughout the literature. The invariances on translation, scale and rotation and the types of SBIR are also analyzed. The complete summary is visualized in Table 1, and we discuss each aspect in the sections below. In Table 1, we treat papers describing the same work but differing in versions (conference or journal) as one work, so the reviewed 22 papers are categorized into 16 works. This categorization will be used from now on. In addition, the discussions in this section are from a statistical perspective and the corresponding technical perspective discussions can be found in Sect. 4.

3.1 Visual cue imbalance

As it is quite intuitive that similar representations will be easier to compare with, out of 16 works, 15 works have addressed the ***visual cue imbalance***, mostly with edge detection. Exceptionally, Sangkloy et al. [47] argued that edge detection is not necessary for deep learning, yet more comparing experiments are desired to verify the point. In general, we can conclude that it is beneficial to address the ***visual cue imbalance*** with edge detection. More technical discussion can be found in Sect. 4.2.

3.2 Content imbalance

For the ***content imbalance***, just a few works (5 out of 16) that studied the sketching-the-object type have considered to address it with either manual [13,17,18] or automatic approaches [29,36,44]. We have to admit that whether to address ***content imbalance*** depends on the concrete application and dataset. If the application is sketching-the-object type and the dataset possesses obviously cluttered background, it is then quite necessary to address the ***content imbalance***. Both [18,29] have shown improvement by addressing the ***content imbalance*** with comparing experiments and they both include sketching-the-object type and employ a dataset with cluttered background. More technical discussion can be found in Sect. 4.1.

3.3 Abstraction gap

For the ***abstraction gap***, the random distortion is the best addressed (15 out of 16) subaspect, since it is intrinsically edge distortion and many techniques in the field can tolerate it. The different ways that different feature types use to tackle the random distortion are summarized in detail in Sect. 4.3.

Simplification is also widely addressed (11 out of 16) by controlling the edge amount and/or the abstraction level of the edge maps. The detailed discussions regarding different features on this subaspect are also included in Sect. 4.3.

Only the unrealistic disproportion is addressed by a single instance [13–15] with a costly deformation process (elastic matching). The unrealistic disproportion is the most challenging among the three subaspects in nature due to the high-level human perception involved. Probably, only part analysis-based approaches could address it but the computational cost would then be higher than holistic approaches.

Alternatively, semantic interpretations of the visual information, i.e., attributes, can be used to cross the ***abstraction gap***, like being demonstrated in [53].

3.4 Translation, rotation and scaling

For the invariances on different properties, the translation invariance has received the most attention (14 out of 16). The scale invariance normally comes with an extra cost and users would tend to draw objects at the proper scale. It is only addressed when it is handy for the employed technique (5 out of 16). As to the rotation invariance, the SBIR system's designing goal would matter, i.e., whether it is desired to have rotation invariance. From the literature, it would be clear to see that most of the works (13 out of 16) chose to consider the rotation as a kind of information that the sketches convey and thus not to address it.

3.5 SBIR types

For the SBIR type, slightly more works (9 vs. 8) studied the sketching-the-scene type while most recent works have focused on the sketching-the-object type due to its commercial value.

3.6 Conclusions

When designing an SBIR system, clearer designing goals would guide to a more effective system. The aspects listed in this section are the ones we highly recommend the authors to consider. Among them, we recommend to consider the SBIR type first, as it will decide if it is necessary to address ***content imbalance***. Anyhow, from the history review, we can see sketching-the-object type is becoming more and more important, since commercial product search is estimated to be a good scenario to apply SBIR. Among all the aspects, ***visual cue imbalance***, ***content imbalance***, random distortion and simplification of the ***abstraction gap***, and invariance on translation are the most important ones to involve and solve with existing techniques. Invariances on scale and rotation are relatively less important and depend on the specific scenario to choose the coping strategies. Unrealistic disproportion is very challenging to solve and would desire future exploration.

4 Summary of the general SBIR framework

The general SBIR framework is illustrated in Fig. 2. As can be seen, the whole framework is divided into the population phase for gallery preparation and the retrieval phase. Two modules are optional: the object localization module and the gallery indexing module. This section summarizes the technologies involved in each module through the eras.

4.1 Object localization

Object localization is an optional module in the population phase. It is very necessary for the sketching-the-object type and for addressing the ***content imbalance***. Initiated in the second era and proliferating in the third era, both manual [13–15] and automatic [17,18,29,36,44] approaches had been applied for this module. Manually drawn bounding boxes [13–15] are very expensive to obtain and thus are hard to apply to real-life applications. Sketch specific masks [17,18] are easy to scale up, but are not so accurate, since the salient object in the image may not be at the same location as in the sketch. On the other hand, hierarchical segmentation [29], prominent edge extraction [44] and object detection [36] are more accurate methods which can more reliably extract the main object of the image out. To scale up, any automatic saliency detection method can be used for this module. Given the large amount of images usually present in the gallery, this module is time consuming in terms of computation. Therefore, it is natural to treat it as an offline procedure applied before the online retrieval.

4.2 Edge extraction

Edge extraction is the mandatory module for addressing the ***visual cue imbalance***. It is also one of the strategies for addressing the simplification subaspect of the ***abstraction gap***. The common techniques for edge extraction are the popular edge detectors like Canny [6] and Sobel [52]. The thresholds in the edge detectors are usually used to control the amount of the edges and thus address the simplification subaspect. Since sketches are generally composed of strokes which are essentially edges, it is intuitive to compare

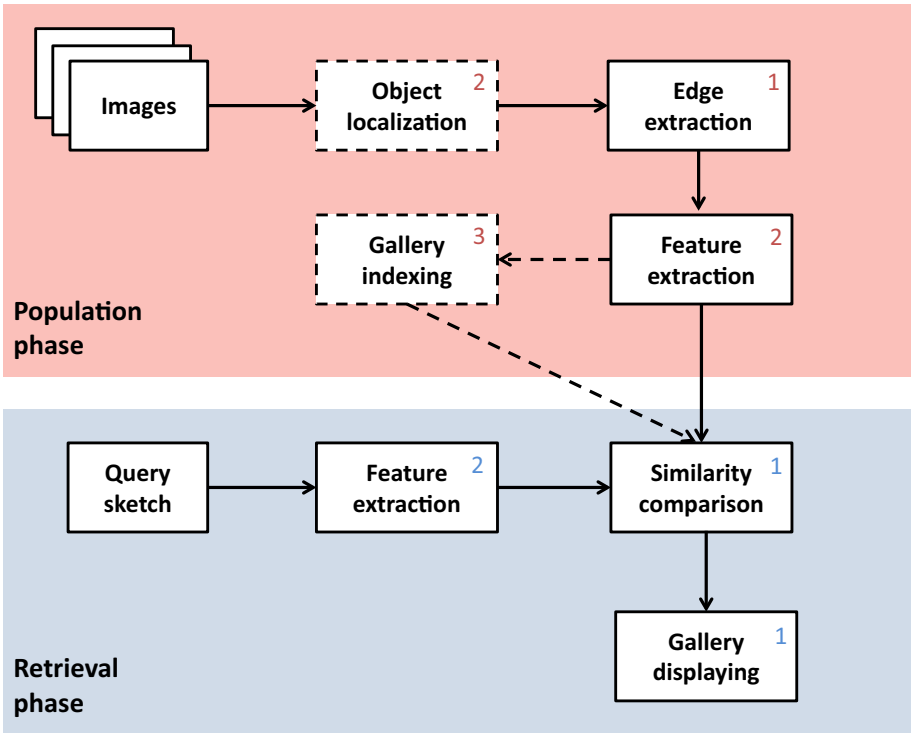

Fig. 2 The general SBIR framework. The numbers in the right-top corner indicate in which era the module was proposed and the boxes with dashed boundaries indicate optional modules

edge maps of the images with sketches. Thus, this module is adopted by almost all the works in the literature.

4.3 Feature extraction and similarity comparison

Feature extraction is the mandatory module for both the population phase and the retrieval phase. It encodes the sketches and the feature maps into representations which are efficient for similarity comparison. During the encoding procedure, it has the potential to address the random distortion and simplification subaspects of the ***abstraction gap***.

This section summarizes the features used in the literature and analyzes their characteristics. We divide the features into 4 types according to their similar feature extraction procedure as : pixel-based, contour-based, histograms and deep features. We summarize each type as below and list the type of feature each work had employed in Table 2. (More detailed description for each work can be found in Sect. 2.) Also, since similarity comparison is closely related to each feature type, we also summarize the similarity comparison methods here together with each feature type and list them in Table 3. The 4 types of features are :

– *Pixel-based* There is no feature extraction step for the pixel-based features where raw pixels are used directly for similarity comparison. The similarity comparison is directly accumulating the pixel value differences. In general, this type of feature is quite rigid and has low tolerance to random distortion and translation. Apart from the first era, few works have adopted this type of feature.
– *Contour-based* The contour-based features try to extract, regulate and encode the salient edge segments of the objects. The contour-based similarity comparison will first find matched segment pairs across two objects, and then accumulate each pair's similarity score to calculate the overall similarity between two objects. The similarity metric between a pair of segments highly depends on the specific representation of the segments. For example, in Chans et al. [9], first degree implicit polynomials (lines) were used as representations. Therefore, the length and the angles of inclination were used for similarity comparison. The contour-based features can deal with random distortion and simplification in the salient segments extraction, but the major drawbacks for this type of features are the high computational cost and the loss of object details. Therefore, only a pair of works have used this type of feature.
– *Histograms* Histogram-based features render some specific aspect(s) of the statistics of the image into a histogram, and use it as the representation for the image. Every image has a histogram representation of the same length. Therefore, by employing metrics for the vectors of the same length, this representation is very efficient for comparing the similarity between different images. Usually, BOW representation is employed together with histogram features for more efficient computation. Several distance metrics can be used for similarity compari-

Table 2 The summary of the types of features employed for SBIR

Year	Author	Pixel-based	Contour-based	Histograms	Deep features
1992	Hirata and Kato [26] ([31])	✓			
1994	Faloutsos et al. [21] ([42])	✓			
1997	Del Bimbo and Pala [13] ([14,15])	✓			
1997	Chans et al. [9]		✓		
2000	Rajendran and Chang [45]			Curvature-direction representation	
2005	Chalechale et al. [8]			APAI	
2010	Eitz et al. [18] ([17])			HOG, Tensor, ARP and EHD	
2011	Eitz et al. [19]			SC, spark feature, HOG and SHOG	
2011	Hu et al. [29]			GF-HOG, SIFT, SSIM	
2011	Cao et al. [7]	✓			
2013	Hu and Collomosse [28] ([27])			GF-HOG, HOG, SIFT, SSIM, SC and Tensor	
2014	Parui and Mittal [44]		✓		
2014	Li et al. [36]			HOG	
2016	Yu et al. [58]				Triplet network, Sketch-a-Net
2016	Sangkloy et al. [47]				Triplet network, GoogLeNet
2016	Song et al. [53]				Triplet network, Sketch-a-Net
No.	16	4	2	7	3

Table 3 The summary of similarity comparison metrics for different feature types

Feature type	Similarity metrics
Pixel-based	Accumulated pixel value differences
Contour-based	Accumulated scores of the matched segment pairs
Histograms	City block, cosine, Chi-square and histogram intersection distances
Deep features	Euclidean distance

son of histograms and a good demonstration for this can be found in Hu and Collomosse [28], where city block, cosine, Chi-square and histogram intersection distances are thoroughly compared. The histogram features can tolerate random distortion due to the grid division scheme in feature extraction. When combined with BOW representation, translation is also tolerated. Due to their efficacy and efficiency, they are most widely used in the literature.

– *Deep features* Deep features also have a fixed length (normally the output of the last feature layer is used) and are easy to compare with each other, e.g., with Euclidean distance [58]. The convolutional layers and max-pooling layers can handle random distortion and translation by design and the max-pooling layers can cope with the simplification as well. But CNN needs huge amount of data to train due to the large set of parameters in it, and this often brings a challenge to the sketch field as still limited quantities of sketches are available.

Table 2 lists the type of feature employed by each work. For the histograms type and deep features type, we also list the exact feature(s)/network architectures for comparison. For the other two types, we do not list the exact details, as the features in them are generally pixels or lines. Through the table, we can observe that it started from the pixel-based features and turned majorly into histogram features, with only a pair of works employing contour-based features. Recently the trend has already begun to favor the deep features. Given limited computational resources, histogram features are the best solution so far, since they do not ask for large amount of training data and do not need good GPU for efficient computation. For the deep features, current studies [47,53,58] have generally focused on utilizing or combining existing advanced deep architectures yet neglected the study on convolutional neural network's ability to address the ***abstraction gap***. Future work on designing specific deep architectures (choosing the number of layers, setting the filters' sizes or introducing other novel network components), aiming at addressing the ***abstraction gap***, is highly desired.

4.4 Gallery indexing

This section summarizes the indexing schemes employed in the literature and discusses about what kind of indexing scheme is desirable for SBIR.

The indexing schemes were mostly explored in the third era and can be categorized into 3 major types: hierarchical clusters (k-means tree and k-medoids tree) [17,18], k-d tree [28] and inverted index (including using both visual words and edgels for entries) [7,19]. All of these schemes can reach sub-linear efficiency. Among these schemes, the usage of k-d tree and inverted index is quite standard and does not reflect the focus of SBIR, i.e., object shapes. But the hierarchical clusters scheme organizes the images with similar shapes into clusters. This behavior could offer more coherent searching results and thus brings special interest into SBIR. Currently, indexing schemes are not considered by all the works except a few mentioned above. Yet an image gallery well organized by shape similarity is beneficial for both accuracy and efficiency.

4.5 Gallery displaying

Gallery displaying module displays the retrieval results to the users. The focus of this module is to offer pleasant viewing experience to the users. So normally, the layout of images on each page, the zoom-in functionality and the image launching speed are the key points of pay attention to.

4.6 Conclusions

As we can see, different modules have different abilities on achieving the SBIR objectives. A complete system with each module well tuned normally would be more effective than just focusing on one module. Therefore, apart from keeping the SBIR objectives in mind, it is also highly recommended to keep all the modules of the general SBIR framework in mind and put effort onto each of them.

5 Summary on datasets and evaluation metrics

To evaluate a SBIR system's performance, the dataset and evaluation metric are quite essential. Here, we look through the existing datasets and evaluation metrics to find out what kinds of datasets and evaluations are more suitable for SBIR.

Table 4 The comparison of the datasets

Year	Publication	Sketch no.	Image no.	Sketch type
1992	Hirata and Kato [26]([31])	18	205	–
1994	Faloutsos et al. [21]([42])	1	1000	–
1997	Del Bimbo et al. [13]	5	100	–
1997	Chans et al. [9]	112	137	Observed sketches and memory sketches*[1]
2000	Rajendran and Chang [45]	–	5000	–
2005	Chalechale et al. [8]	400	4000	–
2010	Eitz et al. [18]	86	1.5M	Traced sketches*[2] and memory sketches
2010	Hu et al. [27] ([29]) (*avail.*)	(i)25*[3] (ii)7	(i)160 (ii)383	–
2011	Cao et al. [7]	132	2.1M	Memory sketches
2011	Eitz et al. [19] (*avail.*)	31	1240 and 100,000*[4]	Imaginary sketches*[5] and traced sketches
2013	Hu et al. [28] (*avail.*)	330	15,000	Memory sketches
2014	Parui and Mittal [44]	(i)175 (ii)7	(i)1.2M (ii)383	–
2014	Li et al. [36] (*avail.*)	84	840	–
2016	Yu et al. [58] (*avail.*)	716	716	Memory sketches
2016	Sangkloy et al. [47] (*avail.*)	75,471	12,500	Memory sketches

The available (*avail.*) datasets are labeled out. If the cell is empty, it means that no reliable source can be identified in the paper
*[1] Observed sketches are drawn while observing the original images, and memory sketches are drawn after careful observation and then without images
*[2] Traced sketches are generated by tracing major contours on top of the images
*[3] Two datasets were employed
*[4] Each sketch has 40 corresponding images and there are 10,000 distractor images
*[5] Imaginary sketches are drawn entirely without images but with imagination

5.1 Datasets

This section lists the datasets employed in the literature and tries to recommend the suitable benchmarks and describe the desired characteristics that could be considered when generating the future new datasets. A comprehensive summary of the datasets is offered in Table 4. From Table 4, we can observe:

- The amount of the images in the gallery was increasing and reached the peak in the third era, from hundreds to thousands until millions. Yet it dropped back dramatically in the 4th era, as people's attention started to focus on fine-grained differences. The amount of the sketches had kept its moderate scale until the 4th era, when the recent works [47,58] started to collect more and more sketches to train the deep networks. Another reason for the increase of the number of sketches is the popularity of some crowdsourcing Internet marketplaces, like Amazon Mechanical Turk (AMT), which make the generation of hundreds of thousands of sketches manageable.
- The sketches can be generated while observing (observed sketches), by tracing (traced sketches), from memory (memory sketches) or by imagination (imaginary sketches). Among these types of sketch generation, the memory sketch is considered the closest type to the real application scenario (i.e., people are trying to search some vague memories with sketches) and has received the most research attentions [7,9,18,28,47,58].

For the benchmarking purpose, 3 recently proposed datasets [36,47,58] are recommended to refer to. They all focus on the fine-grained SBIR concept and quite challenging. Higher performances on them are appreciated. As for designing new SBIR datasets, it would be valuable to hold on the fine-grained concept as it has highlighted the pain point of SBIR as an industrial application : using sketches to extract the shape details and other means like texts to retrieve category. The major challenge for collecting a fine-grained SBIR dataset is the massive supervision needed to rate the similarity between each sketch–image pair. All of Li et al. [36], Yu et al. [58] and Sangkloy et al. [47] have offered different strategies which are good initiatives yet still have limitations. How to obtain this kind of supervision at large scale and with good quality is the aspect worth improving.

5.2 Sketch–image relevance definitions

Before discussing about evaluation metrics, one thing needs to be clarified, that is how to define an image is relevant to a query sketch. We name it sketch–image relevance definition. In the literature, there are 3 types of relevance definitions:

Table 5 The comparison of several evaluation metrics

Year	Publications	Evaluation metric	Type	Relevance
1992	Hirata and Kato [26] (Kato et al. [31])	Hit rate at K*[1] (named recall ratio in the papers)	1	S
1994	Faloutsos et al. [21]	AVRR, IAVRR*[2]	3	M
1997	Del Bimbo and Pala [13]	The correlation between human rankings and algorithm rankings	4	M
1997	Chans et al. [9]	The rank of each reference image	3	S
2005	Chalechale et al. [8]	ANMRR*[3]	3	M
2010	Eitz et al. [18]	Median retrieval rank of the retrieved reference image over multiple queries	3	S
2010	Hu et al. [27] ([28,29])	Average precision (AP)	5	C
2011	Cao et al. [7]	Hit rate at K, precision at K*[4]	1,2	S,M
2011	Eitz et al. [19]	Kendall's tau*[5]	4	M
2014	Parui and Mittal [44]	Precision at K	2	C
2014	Li et al. [36]	Summed similarity scores	4	M
2016	Yu et al. [58]	Hit rate at K, Triplet ranking*[6]	1,4	S,M
2016	Sangkloy et al. [47]	Hit rate at K	1	S
2016	Song et al. [53]	Hit rate at K, Triplet ranking	1,4	S,M

For the 'Relevance' column, 'S' stands for single reference image, 'M' for multiple reference images and 'C' for category relevance
*[1] Over N queries, the portion that has retrieved the reference image among top K results
*[2] AVRR: the average rank of all relevant, displayed images; IAVRR: the ideal AVRR when all the relevant images are ranked on the top
*[3] ANMRR: the mean of the normalized average rank of all relevant images, and it is the mean of multiple queries
*[4] Among the top K results, the portion that is relevant to the query
*[5] Kendall's tau is a measure for rank correlation [32]
*[6] The percentage of the correctly ranked triplets in the retrieved list. It is an evaluation for multiple objects ranking ability. For more detailed explanation please refer to [58]

- Single reference image: only one relevant image for each query. (7 works)
- Multiple reference images: multiple relevant images for each query. (8 works)
- Category relevance: the images of the same category are counted as relevant for each query. (2 works)

It can be clearly seen that the single reference image and multiple reference images are the most popular definitions for sketch–image relevance. Normally, the single reference image definition comes with the specific image search which assumes that the user wants to search one special instance of image in his/her mind. The multiple reference images definition comes with the general image search which assumes that the user just has a concept in his/her mind and wants to retrieve as many similar images as possible. Ideally, both functionalities (specific and general image search) could be investigated, since the users would be interested in both functionalities in different situations.

5.3 Evaluation metrics

Coming with the proposed various datasets, diverse evaluation metrics are also adopted. Continuing the enumeration, here we list the different evaluation metrics used in the literature in Table 5. The qualitative evaluations are not referred, as they are not informative for the comparison of different systems. Also in the literature, for the same evaluation scheme, different works may name it differently. We unify the notions to make the summary clearer.

Overall, there are 5 basic formats of evaluation metrics existing in Table 5 (explanations are available in the table footnotes):

- *Hit rate at K.* (5 works)
  This metric looks at the single reference image among the top K results. Over multiple queries, it is the portion of the queries which have the reference image among the top K results.
- *Precision at K.* (2 works)
  This metric looks at multiple reference images among the top K results. It is the portion of the reference images among the top K results.
- *The statistics of the ranks of the retrieved reference images.* (4 works)
  This metric looks at multiple reference images, but it does not only focus on the top K results. Instead, it takes every reference image's rank into account. It can be many types of statistics of the reference images' ranks, like mean and median.
- *The correlation between human rankings and algorithm rankings.* (5 works)

This metric looks at multiple reference images. It compares the human supervised rankings with algorithm rankings. Human supervised ranks of images are needed and thus it is labor-intensive.

– *Average precision (AP).* (1 work)
This metric looks at category relevance and multiple reference images. It is calculated from the precision and recall curve of the retrieval result.

From the above list, hit rate at K, the statistics of the ranks of the retrieved reference images and the correlation between human rankings and algorithm rankings are the 3 most popular evaluation schemes. Among these 3 schemes, hit rate at k is designed for single reference image, and the rest 2 are for multiple reference images. The statistics of the ranks of the retrieved reference images was mainly used in the first and second eras, and its simplex nature makes it not capable to reflect the relative rankings of the similar images, i.e., which image is more similar to the query sketch. On the contrary, the correlation between human rankings and algorithm rankings is designed for measuring the relative rankings of the similar images and more ideal to evaluate the SBIR system's fine-grained ranking ability, but it usually requires large amount of human rankings for evaluation. In general, we recommend the researchers first to consider hit rate at K for single reference image evaluation and the correlation between human rankings and algorithm rankings for multiple reference image evaluation.

5.4 Benchmarks

For most of the works in the literature, they are evaluating their method(s) on their own dataset. Just a few works have presented a benchmark comparing with other works, yet still in a small scope. We list all the available benchmarks in this section.

In the first and second eras, Del Bimbo and Pala [13] had a benchmark with its precedent works. Table 6 shows an excerpt of it. There are 3 query sketches of different bottles, 22 relevant images of bottles and 100 images in total as the gallery formed with other dissimilar shapes. Given such a small scale gallery, the SBIR's performance is quite satisfied using the precision at K measurement.

Table 6 Benchmark of the first and second eras

Ranking interval system	1–5	1–10	1–22	1–30	1–40
ETM [13]	5	10	21	22	22
QBIC [42]	4	7	11	12	14
QVE [26]	5	10	20	21	21

It uses precision at K measurement. The value of K is changed from 5 (1–5) to 40 (1–40). Each cell lists the number of relevant images among the top K results. There are 22 relevant images among 100 gallery images. In the paper [13], there are results of 3 query sketches. We list the result of the first query here

Table 7 Benchmark on different features for SBIR

Descriptors	Distance measures	k	mAP
GF-HOG	Histogram intersection	3500	0.1222
HOG	Chi-square	3000	0.1093
SIFT	Chi-square	1000	0.0911
SSIM	Chi-square	500	0.0957
ShapeContext	Chi-square	3500	0.0814
Structure Tensor	Chi-square	500	0.0798

'k' represents the best number of the visual words in the BOW representation

Table 8 Benchmark on fine-grained SBIR

	Sangkloy et al. [47]	Li et al. [36]	SP
Airplane	**27.2**	22	20.33
Bicycle	**21.5**	11.67	13.83
Car	15.8	**18.83**	14.5
Cat	**13.8**	12.17	7.67
Chair	**21.7**	20	20.33
Cow	**19.8**	19.67	14
Dog	**21**	9.5	6.83
Horse	23.2	**31.67**	7.33
Motorbike	13	**22.5**	9
Sheep	**21**	17.67	5
Average	**19.8**	18.57	11.88

SP is the abbreviation for spatial pyramid [35]
The best performance for each category are shown in bold

In the third era, Eitz et al. [17–19] and Hu et al. [27,28] had both published some benchmarks on the similar sets of features yet on their different datasets and experiment settings. A representative table can be found in Table 7 (extracted from [28]). It clearly states that HOG-based features, e.g., GF-HOG and HOG, outperform all the other handcrafted features before the emergence of deep features. However, it also states that on million level datasets SBIR's performance is not so good (generally low mAP values).

From the fourth era, fine-grained SBIR concept has been raised and some new datasets have been published for this purpose. One dataset that several works have been compared on is proposed by Li et al. [36], and Table 8 shows a comparison extracted from [47]. We can see that, for fine-grained SBIR, the ability to rank the images of the same category according to the query sketch is focused on. Also, deep features have demonstrated their strength over the handcrafted features.

Apart from the benchmarks, a few works have reported their system efficiency in terms of retrieval time. They are listed in Table 9. We can see that methods using histograms [17,28] can generally finish within several seconds and can

Table 9 The efficiency of different works

Publications	Dataset size	Retrieval time	Remarks
Chalechale et al. [8]	4000	7 s	
Eitz et al. [17]	1.5M	A few seconds	∼0.006 s with *k*-means tree indexing
Cao et al. [7]	2.1M	∼1 s	With edgel indexing
Hu and Collomosse [28]	15k	∼2.5 s	∼0.015 s with *k*-d tree indexing
Parui and Mittal [50]	1.2M	1–5 s	With *k*-medoids tree indexing
Yu et al. [58]	∼100	0.03 s	

be further improved to several milliseconds with a proper indexing scheme. Pixel-based methods [7] and contour-based methods [50] are much slower yet can still satisfy real applications with indexing schemes. Deep features [58] have the similar performance as histograms.

Additionally, we offer some suggestions given the available benchmarks and efficiency reports. Firstly, there is no common benchmark used for most of the works, and we promote the establishment of such a benchmark. Once established, future works can compare with each other more objectively. Secondly, the system efficiency is not well reported. We also promote future works to report the system efficiency together with the system accuracy.

6 Conclusions and future work

Looking through the history of SBIR, many interesting works have published their different strategies for matching sketches and images. However, concrete common objectives and a complete methodology are not raised up to guide the research on SBIR. As a result, the focuses of SBIR researches are frequently shifted and different works are not easy to benchmark together. In this survey, we try to fill in this gap by defining detailed SBIR objectives and a complete methodology covering all the technical aspects. After all the discussions, we conclude as follows:

1. The common objectives are developing strategies of addressing the sketch–image differences, including ***visual cue imbalance***, ***content imbalance*** and ***abstraction gap*** which is further divided into random distortion, simplification and unrealistic disproportion. Edge detection is good for solving the ***visual cue imbalance***. Saliency detection is useful for the ***content imbalance*** and would better be done offline as preprocessing. The ***abstraction gap*** is the most challenging aspect and should be treated as the core objective. Among the subaspects of the ***abstraction gap***, random distortion and simplification have existing solutions but unrealistic distortion is quite difficult to address. Semantic approach like attributes is another path to go through the ***abstraction gap*** yet may be less accurate in shape matching. Therefore, the most important focus for the future works would be improving the ability to address the three subaspects of the ***abstraction gap*** and in other works understanding human abstraction better.
2. The complete methodology is the general SBIR framework. Different modules of the SBIR framework have their distinct usages and some modules can work together to better solve some issues. For example, we can use the gallery indexing to better organize the gallery into clusters with similar shapes, and afterward feature extraction and similarity comparison can find the similar clusters first very efficiently and rank the images inside these clusters according to the query. Nevertheless, only by using all the modules wisely, could an optimal system be built. So, it is recommended to focus on the framework rather than one or two modules in research.
3. For the future research, fine-grained SBIR is the recommended direction. In nowadays searching engines, texts have already been very convenient to retrieve category level information. It is the intra-category fine-grained shape differences that are challenging to differentiate and are a good place to apply SBIR's advantage.
4. To evaluate the system, several public datasets are available as introduced in Sect. 5.1. For evaluating the fine-grained differences, large amount of properly ranked sketch–image pairs are important and hard to obtain. Thus, it is a good problem to address. The correlation between human rankings and algorithm rankings for multiple reference image evaluation would be the important metric for fine-grained evaluation.

Above all, we hope this effort has provided a clearer picture for the SBIR field, and we also hope it has brought more consensuses and enough inspirations.

References

1. Abdulbaqi, H.A., Sulong, G., Hashem, S.H.: A sketch based image retrieval: a rreview of literature. J. Theor. Appl. Inf. Technol. **63**(1), 158–167 (2014)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
3. Bentley, J.L.: Multidimensional binary search trees used for associative searching. Commun. ACM **18**(9), 509–517 (1975)
4. Birari, D.R., Shinde, J.: Survey on sketch based image retrieval. Int. J. Adv. Res. Comput. Commun. Eng. **4**(12), 513–516 (2015)
5. Candemir, S., Borovikov, E., Santosh, K.C., Antani, S.K., Thoma, G.R.: Rsilc: rotation- and scale-invariant, line-based color-aware descriptor. Image Vis. Comput. **42**, 1–12 (2015)
6. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**(6), 679–698 (1986)
7. Cao, Y., Wang, C., Zhang, L., Zhang, L.: Edgel index for large-scale sketch-based image search. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–768 (2011)
8. Chalechale, A., Naghdy, G., Mertins, A.: Sketch-based image matching using angular partitioning. IEEE Trans. Syst. Man Cybern. **35**(1), 28–41 (2005)
9. Chans, Y., Lei, Z., Lopresti, D.P., Kung, S.Y.: A feature-based approach for image retrieval by sketch. In: SPIE International Symposium on Voice, Video and Data Communications, pp. 220–231 (1997)
10. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition, pp. 539–546 (2005)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
12. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 5:1–5:60 (2008)
13. Del Bimbo, A., Pala, P.: Visual image retrieval by elastic matching of user sketches. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 121–132 (1997)
14. Del Bimbo, A., Pala, P., Santini, S.: Visual image retrieval by elastic deformation of object sketches. In: IEEE Symposium on Visual Languages, pp. 216–223 (1994)
15. Del Bimbo, A., Pala, P., Santini, S.: Image retrieval by elastic matching of shapes and image patterns. In: IEEE International Conference on Multimedia Computing and Systems, pp. 215–218 (1996)
16. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. Graph. (Proceedings of SIGGRAPH) **31**(4), 44:1–44:10 (2012)
17. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: A descriptor for large scale image retrieval based on sketched feature lines. In: Eurographics Symposium on Sketch-Based Interfaces and Modeling, pp. 29–36 (2009)
18. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: An evaluation of descriptors for large-scale image retrieval from sketched feature lines. Comput. Graph. **34**(5), 482–498 (2010)
19. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Trans. Vis. Comput. Graph. **17**(11), 1624–1636 (2011)
20. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
21. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. J. Intell. Inf. Syst. **3**(3–4), 231–262 (1994)
22. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
23. Fukunage, K., Narendra, P.M.: A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans. Comput. **24**(7), 750–753 (1975)
24. Gaidhani, P.A., Bagal, S.: Survey paper on sketch based and content based image retrieval. Int. J. Sci. Res. **4**(12) (2015)
25. Gao, Y., Leung, M.K.: Face recognition using line edge map. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 764–779 (2002)
26. Hirata, K., Kato, T.: Query by visual example—content based image retrieval. In: International Conference on Extending Database Technology: Advances in Database Technology, pp. 56–71 (1992)
27. Hu, R., Barnard, M., Collomosse, J.: Gradient field descriptor for sketch based retrieval and localization. In: IEEE International Conference on Image Processing, pp. 1025–1028 (2010)
28. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comput. Vis. Image Underst. **117**, 790–806 (2013)
29. Hu, R., Wang, T., Collomosse, J.: A bag-of-regions approach to sketch based image retrieval. In: IEEE International Conference on Image Processing, pp. 3661–3664 (2011)
30. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142 (2002)
31. Kato, T., Kurita, T., Otsu, N., Kyoji, H.: A sketch retrieval method for full color image database-query by visual example. In: IAPR International Conference on Computer Vision and Applications, pp. 530–533 (1992)
32. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
33. Knutsson, H.: Representing local structure using tensors. In: Scandinavian Conference on Image Analysis, pp. 244–251 (1989)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
35. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
36. Li, Y., Hospedales, T.M., Song, Y.Z., Gong, S.: Fine-grained sketch-based image retrieval by matching deformable part models. In: In British Machine Vision Conference (BMVC) (2014)
37. Li, Y., Song, Y.Z., Hospedales, T., Gong, S.: Free-hand sketch synthesis with deformable stroke models. Int. J. Comput. Vis. **122**(1), 169–190 (2017)
38. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, 20–25 September, 1999, Kerkyra, Corfu, Greece, Proceedings, vol. 2, pp. 1150–1157 (1999)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
40. Manjunath, B.S., Salembier, P., Sikora, t.: Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, New York (2002)
41. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: IEEE International Conference on Computer Vision Theory and Applications, pp. 331–340 (2009)
42. Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G.: The qbic project: querying images by content, using color, texture, and shape. In:

Storage and Retrieval for Image and Video Databases (SPIE), pp. 173–187 (1993)
43. Opelt, A., Pinz, A., Zisserman, A.: Learning an alphabet of shape and appearance for multi-class object detection. Int. J. Comput. Vis. **80**(1), 16–44 (2008)
44. Parui, S., Mittal, A.: Similarity-invariant sketch-based image retrieval in large databases. In: European Conference on Computer Vision, pp. 398–414 (2014)
45. Rajendran, R.K., Chang, S.F.: Image retrieval with sketches and compositions. In: IEEE International Conference on Multimedia and Expo, pp. 717–720 (2000)
46. Rui, Y., Huang, T.S., Chang, S.F.: Image retrieval: current techniques, promising directions, and open issues. J. Vis. Commun. Image Represent. **10**(1), 39–62 (1999)
47. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: Learning to retrieve badly drawn bunnies. In: SIGGRAPH (2016)
48. Santosh, K., Lamiroy, B., Wendling, L.: DTW-radon-based shape descriptor for pattern recognition. Int. J. Pattern Recogn. Artif. Intell. **27**(3) (2013)
49. Santosh, K.C., Lamiroy, B., Wendling, L.: Symbol recognition using spatial relations. Pattern Recogn. Lett. **33**(3), 331–341 (2012)
50. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
51. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
52. Sobel, I., Feldman, G.: An Isotropic 3x3 Image Gradient Operator for Image Processing. In: Pattern Classification and Scene Analysis, pp. 271–272 (1973)
53. Song, J., Song, Y., Xiang, T., Hospedales, T., Ruan, X.: Deep multi-task attribute-based ranking for fine-grained sketch-based image retrieval. In: British Machine Vision Conference (2016)
54. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
55. Tang, X., Wang, X.: Face sketch recognition. IEEE Trans. Circ. Syst. Video Technol. **14**(1), 50–57 (2004)
56. Thayananthan, A., Stenger, B., Torr, P.H.S., Cipolla, R.: Shape context and chamfer matching in cluttered scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2003)
57. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)
58. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
59. Yu, Q., Yang, Y., Song, Y., Xiang, T., Hospedales, T.: Sketch-a-net that beats humans. In: British Machine Vision Conference, pp. 7.1–7.12 (2015)
60. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Comput. Surv. **38**(2), 6.1–6.56 (2006)

Yi Li got his PhD degree from Queen Mary University of London. He specializes in the research of sketch recognition, sketch-based image retrieval and sketch synthesis. Now he is working as a post-doc researcher in OSRAM Corporate Technology.

Wenzhao Li is studying in Queen Mary University of London. She focuses on improving graph matching techniques and applying them to various applications. Sketch-based image retrieval is one of the interesting fields she is exploring.