J. Vis. Commun. Image R. 32 (2015) 20–54

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Semantic content-based image retrieval: A comprehensive study ☆

Ahmad Alzu'bi [a,*], Abbes Amira [a,b], Naeem Ramzan [a]

[a] School of Engineering and Computing, University of the West of Scotland, Paisley, PA1 2BE, United Kingdom
[b] College of Engineering, Qatar University, Qatar

ARTICLE INFO

Article history:
Received 8 February 2015
Accepted 13 July 2015
Available online 22 July 2015

Keywords:
CBIR
Image features
Dimensionality reduction
Deep learning
Relevance feedback
Image annotation
Visualization
Semantic gap

ABSTRACT

The complexity of multimedia contents is significantly increasing in the current digital world. This yields an exigent demand for developing highly effective retrieval systems to satisfy human needs. Recently, extensive research efforts have been presented and conducted in the field of content-based image retrieval (CBIR). The majority of these efforts have been concentrated on reducing the semantic gap that exists between low-level image features represented by digital machines and the profusion of high-level human perception used to perceive images. Based on the growing research in the recent years, this paper provides a comprehensive review on the state-of-the-art in the field of CBIR. Additionally, this study presents a detailed overview of the CBIR framework and improvements achieved; including image preprocessing, feature extraction and indexing, system learning, benchmarking datasets, similarity matching, relevance feedback, performance evaluation, and visualization. Finally, promising research trends, challenges, and our insights are provided to inspire further research efforts.

1. Introduction

Different forms of multimedia resources (i.e. text, images, audio, and videos) are rapidly growing with a huge development of visual contents and technologies, e.g. data visualized in smart phones, 2D/3D applications, web content, telecommunication, etc. The new century has witnessed an unparalleled evolution in the amount, availability, complexity, diversity and importance of images in all domains. Therefore, the demand for image services becomes more imperative. Images play a crucial role in a wide range of applications and fields such as education, medical care, weather forecasting, criminal investigation, journalism, advertising, art designs, web, social media, and entertainment. However, visual media require a considerable amount of processing and storage, which requires a highly efficient method to index, store, analyze, and retrieve visual information from image databases. Moreover, images have a little metadata by which they can be indexed and searched conveniently. Consequently, searching images rapidly, accurately, and efficiently for all types of image datasets becomes one of the most challenging tasks.

Humans can properly describe and interpret image contents, including its overall topology and objects using high-level semantic concepts. Unlike humans, digital machines can provide fewer semantic words for the same image. Since machines deal with low-level features extracted from image pixels, it provides a numerical description of images but with a wide gap compared to the human interpretation of the same image. This gap between the richness of high-level human's perception and low-level machine's descriptions is known as the 'semantic gap'. The user looks up for semantic similarity, but the database can only provide similar images by a digital processing. In addition, the semantic gap between image properties and object properties broadly limits the retrieval efficiency [1].

The majority of popular search engines provide image retrieval services using the traditional text-based approach (i.e. captions), and facilitate a self-broadcasting option which causes problems associated with inaccuracies. When a user submits a textual query seeking for certain images, the search engine generally matches keywords in the user's query with indexed terms associated with images in the database. Such approach returns some irrelevant results due to some inappropriate and unrelated terms that used for image indexing. Since textual-based image retrieval methods mainly depend on the metadata associated with images, it requires a manual image annotation by humans. However, manual image annotation or tagging is an imprecise and time consuming process. Moreover, humans have different interpretations for the same image due to the disparity of their knowledge, experience, intelligence, culture background, visual analysis, synonyms, and labeling skills. An automatic annotation by machines could speed up the process, but it still depends on the accuracy of detecting image's edges, colors, object's state, and spatial relations between objects.

☆ This paper has been recommended for acceptance by Prof. Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: ahmad.alzubi@uws.ac.uk (A. Alzu'bi), abbes.amira@uws.ac.uk (A. Amira), naeem.ramzan@uws.ac.uk (N. Ramzan).

http://dx.doi.org/10.1016/j.jvcir.2015.07.012
1047-3203/

Though valuable achievements of prolonged research efforts in this context, the reliance on text-based image retrieval does not satisfy user needs. Therefore, the need arose for alternative retrieval approaches which visually analyze the image contents. Content-based image retrieval (CBIR) is a popular technique that has been widely applied to address the problems of traditional lexical matching systems. In CBIR, images could be retrieved either using low-level features (e.g. color, shape, and texture) that directly extracted from the database, or using high-level features (e.g. events) that also called semantic features. Semantic-based image retrieval mainly matches a user query based on some perceptual contents rather than similarities between captions.

Many systems have been proposed and implemented using CBIR technologies, e.g., query by image content (QBIC) [2], VisualSeek [3], SIMPLicity [4], and Blobworld [5], Such these systems have concentrated on extracting low-level features from image (e.g. color, texture, and shape) in order to represent its semantics. Other systems such as WebSeek [6] and Image Rover [7] have applied the image search based on submitted query keywords, and then performed user classification by providing category browsing and search-by-example facilities. After the early success of these systems, research directions have been extended over the last years. Research efforts in the field of CBIR currently concentrate on profound and challenging problems at different disciplines such as pattern recognition, machine learning, computer vision, and artificial intelligence.

As aforementioned, the end-user targets retrieving the most relevant images by using different forms of query captions minimal interactions with the system. In CBIR, the user interaction is indispensable because users are usually involved in a relevance feedback process to revise the retrieved results as a list of ranked images. Extensive experiments on many CBIR systems showed that low-level image descriptors predominantly fail to describe the high-level semantics in the user's mind [8]. In addition, the CBIR is mainly based on the extraction of low-level features, which should be strictly considered while developing any image retrieval system. Generally, feature extraction approaches attempt to generate a discriminative representation for original images which is one of the key challenges in the retrieval process. Recently, low-level image features have been fused with textual features (i.e. keywords or tags) in many CBIR systems to represent images semantically (i.e. annotation-based retrieval). The automatic annotation of images significantly increases the retrieval accuracy and has attracted more attention recently. Additionally, images can be segmented into a set of regions to represent image objects and every region is treated as a separate image (i.e. region-based image retrieval). Another important factor in the CBIR context is image indexing, especially in high-dimensional data spaces that need to be reduced to lower dimensions for performance requirements (e.g. accuracy, speed, scalability, and memory usage). This problem is known as 'curse of dimensionality', which is very critical in real-world systems (e.g. web search).

The selection of certain similarity matching techniques (i.e. distance metrics) will significantly affect the retrieval accuracy. In image retrieval, a threshold on the minimum acceptable similarity is usually imposed to limit the number and relevancy of retrieved images [9]. In addition, the accuracy of retrieved images can be further improved by enabling system learning in order to develop a pattern model and make predictions or decisions, rather than following only explicitly programmed instructions [10]. Many learning approaches have been conducted and improved over the recent years, which have a substantial impact on the quality of CBIR systems. Upon these interesting issues and challenges, the wheel of research in CBIR has rapidly accelerated towards developing more robust semantic CBIR systems to address these problems; especially the semantic gap. As expected, a wide range of research efforts has been made on several important issues such as visual image features and their combination, effective high-dimensionality reduction and indexing, classification and clustering, similarity matching, relevance feedback, learning models, image datasets, visualization and performance evaluation. Therefore, CBIR systems should be continually reviewed and assessed to exploit their benefits and inspire more constructive efforts and advances.

Some studies and surveys have been presented in the CBIR literature to address and review the most challenging issues and state-of-the-art. The working conditions of content-based retrieval, including patterns of use, types of pictures, role of semantics, and sensory gap are discussed in [1]. Their survey has also presented some aspects of system engineering such as databases, system architecture, and evaluation. An overview provided in [11] on the functionality of various image retrieval systems in terms of technical aspects such as querying, feature extraction, matching, data indexing, and result presentation. They have technically compared specific systems rather than general architectures. Jorgensen [12] has presented a thorough study in the field of image retrieval from many different disciplines and perspectives. Ritendra et al. [13] also have discussed some key challenges involved in the adaptation of image retrieval techniques to build useful systems that can handle real-world data. Mechael et al. [14] have reviewed many articles developed for content-based multimedia information retrieval and discussed their role in current research directions and challenges. Liu et al. [15] have presented a survey on different aspects in the CBIR area with more emphasis on region-based image retrieval, including low-level image feature extraction, similarity measurement, and deriving high-level semantic features. Datta et al. [16] have introduced a comprehensive review on the state-of-the-art of image retrieval. Priyatharshini et al. [17] have focused on different methods that combine visual and textual cues as association-based image retrieval and their evaluation techniques used in the CBIR area. Other studies have been presented on exceedingly related aspects such as relevance feedback [18], high-dimensional indexing [19], and image retrieval learning [20].

Over the last few years, CBIR methods have tremendously grown not only in the research volume, but also in the number and intricacy of the new explored directions. This paper provides a comprehensive study on different aspects and techniques involved in the field of CBIR, including the framework structure, image preprocessing, extraction of visual and textual features, high-dimensionality reduction and indexing, learning approaches in the retrieval process, relevance feedback, similarity matching, visualization and performance evaluation metrics. Recently, some of interesting and promising techniques have been presented in the field which is expected to make a breakthrough in image retrieval domain. For instance, deep neural networks have attracted a lot of research interests in computer vision community. These networks could be trained and utilized as high-level descriptors for image visual contents which may solve unrelated classification tasks [21,22]. Another breakout advancement in the CBIR scope is the successful utilization of local image descriptors which extract distinctive invariant features from images. This provides more reliable matching between different views of objects or scenes. Such features are invariant to image scale and rotation, and robust in matching across a substantial range of affine distortion, addition of noise, change in 3D viewpoint and illumination.

This study is inspired by these recent advancements along with the aforementioned challenges in the field of CBIR. It is essential to review and analyze efforts, challenges and new trends presented, and to provide some thoughts and insights for future research. In particular, this study aims at addressing the following important issues in the CBIR field:

- Are single/combined visual features enough to represent image contents, or should they be fused with additional textual descriptors in order to mimic human semantics and reduce the semantic gap?
- The capability of new proposed methods in reducing the 'curse of dimensionality' problem when compared with traditional methods.
- The ability of CBIR systems to effectively exploit the expected breakthrough by using deep neural networks for learning CBIR tasks.
- How much improvement could be achieved by image annotation methods in order to minimize the user intervention (i.e. user feedback and manual labeling) in this process?
- The exigent need for robust measures to evaluate the performance of image retrieval systems in terms of accuracy, computation complexity, memory usage, and rank among other systems in the CBIR scope.
- How much the achieved advancements can be modeled and propagated to boost real-time retrieval applications; especially over the web?

A comparative analysis is included throughout this study with a view to address the mentioned and other challenging issues. Furthermore, a statistical review of image datasets utilized in CBIR systems is presented. Current and future trends are discussed with the aim of highlighting some contributions and directions, and further inspiring more research efforts in the field of CBIR. The remaining part of this paper is organized as follows: Section 2 extensively presents the components of the CBIR framework, image retrieval methods, and challenges. Section 3 provides a detailed review of image datasets used for benchmarking and evaluation. Performance evaluation measures are discussed in Section 4. Current trends and research challenges with our insights are presented in Section 5. Section 6 concludes this paper. Finally, all acronyms used in this paper are summarized in Appendix A.

2. CBIR framework

This section presents the general framework of CBIR system, and each component of this framework will be illustrated in the following subsections. As shown in Fig. 1, image retrieval system includes a set of correlated blocks and coherent steps.

Firstly, the conventional retrieval scenario begins when a user submits a query image to the system. Both the query image (online process) and all images (offline process) in the database are processed and represented in the same way in order to retrieve only relevant images. Secondly, some preprocessing methods might be applied to the image which mainly depends on the aim of the retrieval application. For instance, the image could be segmented into many smaller blocks or regions that are further processed to represent some image objects. In addition, these smaller parts of the image might be classified or clustered in some categories to be used as region-based retrieval or for learning purposes. Other preprocessing tasks include image resize, rescale, de-noise, etc. Thirdly, visual/textual descriptors are extracted from images and characterized in a certain manner into the data space. Some common extracted features are color, texture, shape and local descriptors. Some techniques apply some preprocessing tasks such as classification or spatial processing after feature extraction, thus the preprocessing of images could be conducted before or after feature extraction. Finally, the system computes the distance between the transformed features of query image and all images in the database in order to return the most relevant images based on some distance measures. The returned images are usually presented as a ranked list. Some retrieval approaches enable users to decide the relevancy degree of retrieved images as a satisfaction measure, i.e. relevance feedback. This may improve the retrieval accuracy by updating the query and similarity measures according to the user's preferences. Automatic feedback and system self-training are preferred for reducing the user intervention and avoiding multiple iterations of refinements.

As an integral part of the CBIR framework, data visualization has recently witnessed a considerable utilization. This addresses the problem of designing graphical user interfaces (GUI) for image representations, query submission and refinement, relevance feedback, and browsing mechanisms. A proper visualization during human–machine interactions guarantees the improvement of retrieval accuracy, maximum flexibility with minimum complexity, and intuitive retrieval environment. This study will also highlight the connotation of visualization in the CBIR domain and its relationship with other CBIR components such as relevance feedback, querying, visual analytics, and image representations.

It is very clear that the image is the core part of processing in the described CBIR framework. Basically, any digital image might

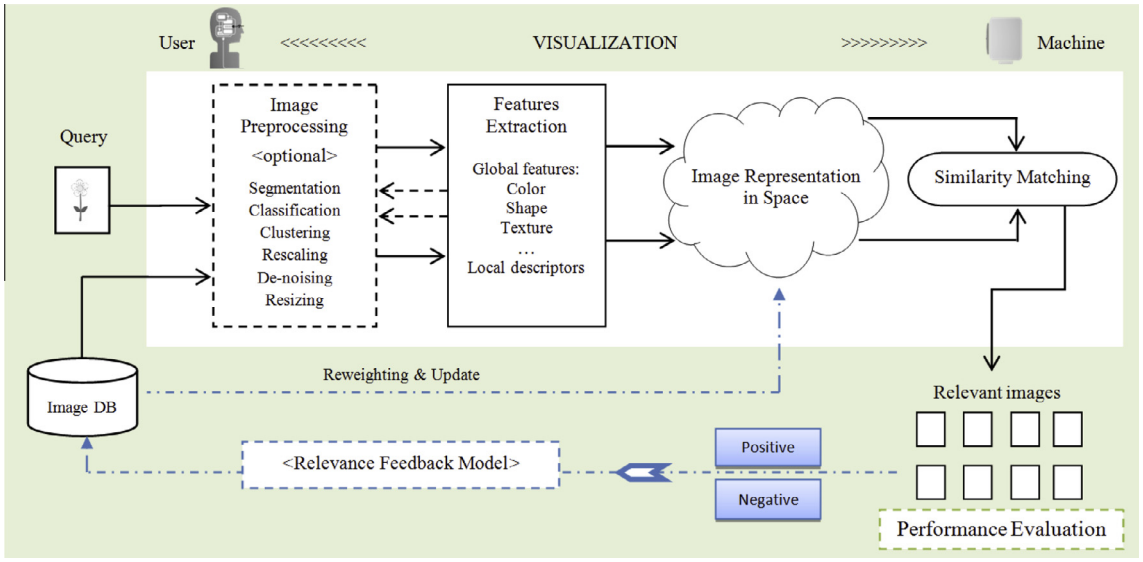

Fig. 1. The general framework of CBIR system.

be characterized by the amplitude of its intensity/grayscale at certain spatial coordinates, and this can be described by two-dimensional function $f(x,y)$ where f is an amplitude at the coordinates x and y. Since images are defined over two dimensions, the digital image processing could be modeled in the form of multidimensional systems. In [23], a general paradigm has been described for image processing where three types of computerized processes are considered: low-level, mid-level, and high-level processes. The low-level processing of images includes primitive operations such as image preprocessing to reduce noise, and contrast enhancement. The mid-level processing of images includes tasks such as segmentation to partition the image into regions or objects, description of those regions and objects, and classification for objects. Finally, the high-level processing of images includes cognitive functions, image analysis, and recognition.

Generally, the low-level processing is characterized by the reality that both inputs and outputs are images. In the mid-level processing, inputs are images, but outputs are the attributes extracted from those images such as object identities and edges. In the CBIR context, images and their attributes are passed for further cognitive analysis in order to return the relevant images to them as outputs. Additionally, image annotation (where the images or segments are described by some text, keywords or tags) generates some textual descriptors and attributes which usually used in the high-level processing. Based on this vision, the field of CBIR encompasses processes that belong to several exceedingly close disciplines such as computer vision, pattern recognition, machine learning, and artificial intelligence. In the following subsections, we will deeply present steps of this framework, adopted techniques, challenges, and new advances with their achievements.

2.1. Image segmentation

Segmentation is one of the most widely applied preprocessing approaches where image pixels are subdivided into some constituent regions or objects that represented by many regions. Region and object detection mainly depends on the application and the homogeneity criteria used. Moreover, the accuracy of segmenting and detecting regions of interest is also influenced by the level at which the segmentation process should stop, i.e. thresholding. Extracting image features from the whole image is simpler than obtaining them from image regions; thus the region-based image retrieval is comparatively considered as closest to the human perception [24]. Therefore, image segmentation is one of the difficult tasks in CBIR systems, especially if it used in further processing stages (e.g. classification and annotation). Since the majority of preprocessing in CBIR domain is segmentation-based and benefits from the semantics obtained from segments/regions, this section will specifically focus on image segmentation. Other preprocessing approaches such as denoising, rescaling, and resizing are problem-dependent and employed as a preliminary processing for optimization and improved results. Accordingly, these preprocessing approaches will be successively discussed in different methods throughout this study.

It is important that segmentation methods should perceptually capture important groupings or regions that often reflect global aspects of the image, and be highly efficient [25]. Based on many studies [25–34] on image segmentation, image segmentation methods are categorized into eight groups as shown in Fig. 2. Among image segmentation techniques, many of successful ones benefit from mapping image elements onto a graph. The segmentation problem is then solved in a spatially discrete space by efficient tools from graph theory. One of the vital advantages of formulating the segmentation on a graph is that it might require no discretization by virtue of purely combinatorial operators and thus incur no discretization errors [26]. Consequently, this section provides more details on graph-based segmentation approaches than other ones. Additionally, graph-based representations and similarities will be successively discussed in this paper.

(1) ***Graph-based segmentation:*** A set of points in an arbitrary feature space is represented as a weighted undirected graph $G(V,E)$, where V is a set of nodes as points in the feature space, and an edge E is formed between every pair of nodes. The weight of each edge, $w(i,j)$, is a function of the similarity between the nodes I and j. Wu et al. [35] have introduced the general method of graph segmentation with a global cost function. After that, many techniques have been proposed such as the graph cut algorithm [36] which is treated as one of the leading methods that has been extended to solve many problems in computer vision applications. Categorizing the graph-based methods is a difficult task. However, they can be broadly grouped, based on some research studies [25,28,33], into the following categories:

(a) *Minimal spanning tree (MST):* It is the spanning tree of weighted and undirected graph such that the sum of weights is minimized, and the intrinsic structure of a dataset can be estimated based on this tree. Clustering the image pixels is performed on the minimal spanning tree. The connection between graph vertices satisfies the minimal sum of the defined edge weights, and the partition of a graph is obtained by removing edges to form various sub-graphs. Some of representative algorithms and recent variants are: exact MST [37], sparse graph [38], dual-tree [39], distributed MST [40], approximate MST [41–43], and recursive MST [44].

(b) *Minimal and normalized graph cut:* Graph cut is the partitioning of graph vertices into two disjoint subsets that are joined by at least one edge, and the minimum cut of a graph is the cut set which has the smallest number of unweighted edges or smallest sum of weights possible. Shi et al. [27] have introduced the normalized cut criterion which measures both the total similarity within groups as well as the total dissimilarity between them. To optimize this criterion, an efficient computational technique based on a generalized eigenvalues problem has been employed. By optimizing these cost functions of graph cut, it becomes possible to get the desirable segmentation. Some of representative techniques and recent variants are: minimum cut [45,46] and normalized cut [27,47–50].

(c) *Graph cut on Markov random field (MRF):* The MRF theory has been introduced as a consistent approach for modeling contextual information such as image pixels and

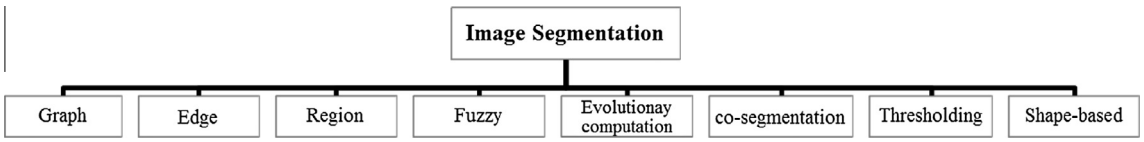

Fig. 2. Image segmentation methods.

visual features. In the context of image segmentation, four sets are considered: S which is a set of all image pixels, N which is a neighborhood defined on that set, L which comprises a set of labels that represent different image segments, and random variables in set X which denote the labeling of image pixels. In particular, each configuration x of the MRF defines a segmentation, and the maximum a posteriori MAP-MRF estimation can be identified as energy minimization problem where the energy corresponding to the configuration x is the negative log likelihood of the joint posterior probability of the MRF [51]. Consequently, the image segmentation problem can be solved by finding the least energy configuration of the MRF. The function optimization is usually obtained by the min-cut/max-flow algorithms. Some representative techniques and recent variants are: s/t graph cut [52], multi-labeling [53], shape prior [54,55], and interactive graph cut [56,57].

(d) *The shortest path:* Given a graph, the object boundary is defined based on a set of shortest paths between pairs of vertices. This approach requires user interactions to handle the segmentation; thus it can provide a friendly feedback with more flexibility [33]. Some representative works on the shortest path approach are in [58,59].

(e) *Local Variation:* Felzenszwalb et al. [60] have described an efficient graph theoretic algorithm for image segmentation. The main idea is to partition the image in such a way that for any pair of regions, the variation across neighboring regions should be larger than the variation within each individual region.

(f) *Eigenvector-based methods:* The input of this method is a matrix of pairwise proximities in a two dimensional image; the output is closely related to the molecular concept of "bond order" matrix which indicates whether any two features do or do not belong to the same cluster. Some of representative algorithms are [27,61,62].

(g) *Random walker:* Leo et al. [63] have proposed this method for performing semi-automated and multi-label image segmentation. Given a number of pixels with user-crafted labels, it could be analytically determined that the probability of starting a random walker at each unlabeled pixel will first reach one of pre-labeled pixels. High-quality image segmentation may be achieved by assigning each pixel to the label for which the greatest probability is calculated. In addition, the theoretical properties of random walker are developed along with the corresponding connections to electrical circuits and discrete potential theory. This algorithm is formulated in a discrete space, such as a graph, based on combinatorial analogs of standard principles and operators from the continuous potential theory; allowing it to be applied in an arbitrary dimension. Some variants of this method have been proposed in the field [26,64,65].

(h) *Dominant set:* Pavan et al. [66,67] have developed a framework for image segmentation based on a new graph-theoretic formulation of clustering. This method is based on the analogies between the axiomatic concept of the cluster and that of a dominant set of vertices, i.e. the notion which generalizes a maximal complete sub-graph to edge-weighted graphs. A correspondence between the extreme of a quadratic form over the standard simplex and dominant sets has been established; thereby it allows using continuous optimization techniques such as the replicator dynamics of the evolutionary game theory.

(2) ***Edge-based segmentation:*** Generally, the process of edge detection is one of the widely used techniques in image processing, which identifies and locates sharp discontinuities in the image. Early edge detection approaches such as Sobel edge detector [68], Canny edge detector [69,70], and Robert edge detector are based on the abrupt changes in image color or intensity. Basically, edges can be formed in gray-level images by using some scalar functions that approximate Laplacians or gradients of images [28]. Many related methods [71–73] have been proposed over the last decade.

(3) ***Region-based segmentation:*** In the region-based segmentation, object pixels are grouped together and marked. The important principles are the spatial proximity which consists of region compactness and Euclidean distance, and the value similarity which includes gray value differences and gray value variances. The region-based segmentation method also requires the use of appropriate thresholding techniques [74]. Region based segmentation methods are broadly categorized into two approaches: split-and-merge and region growing. Split-and-merge methods start with the image as an initial inhomogeneous partition on which a continuous splitting is performed until homogeneous partitions are obtained. Quadtree [75] is one of the commonly used algorithms in this context. After the splitting process, many regions (small fragments) need to be connected in a certain manner. The merging process guarantees that the homogeneity requirements are met and associates the neighboring regions until maximally connected segments can be generated [28]. Some algorithms involved in the split-and-merge segmentation are: region adjacency graph (RAG) [76] and Gaussian Markov random field (GMRF) [77]. Region growing algorithms typically scan the image, in some predetermined manner (e.g. left to right and from top to bottom), in order to compare the current pixel to an already existing but not necessarily completed neighboring segment. If the values of the pixel and segment are close enough then the pixel is added to that segment and this latter is updated. If there are more than one region close enough then the pixel is added to the closest region, but if there is no neighboring region closed enough then a new segment is defined. This process is next reiterated from the next pixel throughout the entire image [78]. Based on the region growing, the watershed transformation [79] is a morphological method defined to segment the image. In this approach, pixel values are considered as topographic data characteristic of a relief, and the value of each pixel denotes the elevation of point. Some of recent algorithms on region-based segmentation have been described by several studies [80,82].

(4) ***Fuzzy-based segmentation:*** Basically, 'chain' is a sequence of pixels where the consecutive pixels are adjacent, and 'link' is a pair of adjacent pixels. In the fuzzy-based segmentation, the strength of links is automatically defined based on statistical properties of the links within regions that identified by the user as belonging to an object of interest. The strength of any chain is the strength of its weakest link. The strength of the strongest chain between any pair of pixels defines the fuzzy connectedness between them. In consequence, the fuzzy object, including a given pixel at a specified threshold, is the set of all pixels whose fuzzy connectedness to the given one is larger than or equal a given threshold [83]. A particular membership function typically assigns such values through a co-domain equal to the closed interval $[0,1]$ of real numbers, where the full membership to a fuzzy set is achieved at the value 1 and a non-membership at the value 0. Some of representative algorithms have been presented in [84–86].

(5) ***Evolutionary computation-based segmentation:*** To address the problem of highly computation cost that resulted during image segmentation, evolutionary computation (EC)-based algorithms are performed. These methods solve the problems associated with large dimensionality spaces using a natural selection which has been shown as a powerful search method. Recently, Yuyu et al. [87] have presented a deep study on EC-based segmentation algorithms. Generally, many segmentation-based works combine the EC approach with other segmentation algorithms such as threshold, region growing, and partial differential equations. In these hybrid methods, the EC takes the role of optimizing parameters or minimizing/maximizing objective functions. Additionally, the EC techniques could be applied to generate segmentation algorithms from a subset of basic image operators such as filters, histogram equalization and threshold [87].

(6) ***Co-segmentation methods:*** Recently, more attention paid on the unsupervised image co-segmentation approach, where the segments are forced to be consistent across a collection of similar images. Many natural image collections contain similar or related objects. For instance, photo collections of a particular theme (e.g., "grazing animals") invariably contain shared contents. In this method, the main goal is to establish some relations across images, and to obtain consistent segmentations that agree with the segmentation clues provided by all of images together. This formulation turns out to perform much better than a single image segmentation method. However, the existing techniques generally have a limitation where the input images must all contain the same set of objects [32]. Some of the recent works on the joint co-segmentation have been discussed in [32,88–91].

(7) ***Thresholding-based segmentation:*** Thresholding is considered as one of the simplest and most commonly used methods for image segmentation. Basically, the image objects, edges, shapes, and backgrounds can be separated by detecting the discontinuities based on a predefined thresholding value. This method possesses the advantages of smaller storage space, fast processing speed and ease in manipulation. Since thresholding is a well-researched field, there exist many algorithms for determining an optimal threshold of the image. However, these methods are sensitive to several transformations such as nonstationary and correlated noise, ambient illumination, and contrast. Thresholding methods have been discussed in [92].

(8) ***Shape-based segmentation:*** Statistical shape models (SSMs) and active contours (ACs) have been widely used to segment objects of interest in images. The most commonly used SSM-based models are the active shape model (ASM) and the active appearance model (AAM). The ASM [93] consists of building a point distribution model from a training set and an iterative searching procedure to locate an instance of such shapes in a new image. The strategies of ASM mainly compose the initialization, matching point detection, and pose and shape parameter update. The ASM models have been extensively utilized in many medical applications such as localization of optic disk's boundaries [94], automated 3D segmentation of lungs [95], and MRI bone segmentation [96]. The AAMs [97] form a framework to statistically model the object shape and local grayscale appearance (e.g. texture variation). These models use the labeled points on image objects that aligned into a common coordinate frame and represented by a vector. The AAM-based segmentation has been employed in many applications such as medical image analysis [98] and face interpretation [99]. In AC models [100], the object in a given image, which has some edges described by a closed curve, is equivalent to the location of sharp image intensity variations by iteratively deforming a curve towards the object's edges [101]. This allows a contour deformation to minimize a given energy functional and produce the desired segmentation. ACs can be broadly categorized into edge-based (e.g. geodesic and geometric) and region-based (e.g. snakes and region growing) models.

In addition to the segmentation methods mentioned, other related approaches have been proposed such as: partial differential equation (PDE) [102], statistical level set [103], clustering-based, and artificial neural network. Since they are used in clustering and classification problems, the latter two methods will be discussed in Section 2.3. Table 1 summarizes the image segmentation methods and more characteristics provided.

2.2. Low-level features extraction

Over the last years, assortments of low-level image descriptors have been proposed in the literature for image representation and indexing. The extracted image features are generally categorized into two types: global and local. Global image features (e.g. color, texture and shape) usually describe the whole image and contain representative information obtained after analyzing image pixels. Local image features specifically describe some parts or key points in the image such as corners and edges that commonly obtained by the segmentation process. In this section, the global and local image features will be illustrated to highlight their importance in CBIR systems, and present the noticeable research interest in recent algorithms especially those based on local image descriptors.

2.2.1. Global image features

The most commonly extracted features in image retrieval systems are color, texture, shape, and spatial locations. This section presents and discusses these features as follows:

2.2.1.1. Color. It is one of the most extensive vision characteristics due to its close relation with image objects, foregrounds, and backgrounds. The color is also a robust visual feature as it does not depend on the state of image contents such as the direction, size and angle. The popular color representations that have mainly been used are color histogram, color moments [104], color correlogram [105], and color co-occurrence matrix [106]. Generally, color spaces are classified into linear color spaces (e.g. RGB, XYZ, CMY, YIQ, and YUV) and non-linear color spaces [107] (e.g. L*a*b, HSV, Nrgb, Nxyz, and L*u*v). The RGB color space is an additive color space based on three primary colors: red, green and blue. Secondary colors can be generated by using the primary colors, for example, using red with blue makes magenta, green and blue makes cyan, and a combination of red, green, and blue at full intensity makes white. However, the RGB space is not very efficient in dealing with real-world images, thus it avoided in most of image retrieval algorithms because it lacks the ability of measuring the perceptual similarity. Moreover, the distances in RGB space have a little meaning in terms of human visual perception. Accordingly, the HSV color space is employed instead of the RGB color space because the components of hue and saturation are very close to human visual perception. The HSV model has three constituent components: The 'hue' which refers the color, the 'saturation' which refers the "vibrancy" of the color, and the 'value' which refers the brightness of the color. As another color space, the YCbCr color space is divided into luminance (Y) and chrominance (Cb, Cr), while Cb and Cr denote the blue–yellow and red–green color difference, respectively. The L*a*b* color space is also derived from

Table 1
Summary of image segmentation characteristics.

Methods	Attributes	Limitations	Refs.	Semantic contributions
Graph-based	– Offer both automatic and user-interactive methods – Provide well-defined relationship between segments	– Unnatural bias toward finding small components – NP-hard to solve – No specific quantitative measure for evaluating the segmentation quality	[25,28,33,35–63,26,64–67]	Partitioning a graph into several sub-graphs such that each of them represents a meaningful object of interest in the image
Edge-based	– Based on the abrupt changes in image intensity or color – Simple and robust on images have enough contrast	– Sensitive to noise – Inefficient on smooth images	[28,68–73]	Similar to how humans segment images
Region-based	An iterative process is performed until some uniformity criteria are satisfied	– Region initialization – Sensitive to a different similarity measures	[74–82]	Provide multiscale low-level hierarchical segmentation, obtaining high quality object candidates
Fuzzy-based	– Noise insensitive – Preserves image details	– Computationally expensive	[83–86]	A fuzzy semantic relates a pair of concepts to a degree of membership
Evolutionary computation-based	Most are based on genetic techniques to deal with parameter optimization or pixel-level problems in image segmentation	– Domain-specific – The evaluation process is computationally expensive	[87]	Construct high-level features to improve the performance of complex image segmentation tasks
Co-segmentation	Obtain consistent segmentations that agree with segmentation clues provided by all the images together	Input images must all contain the same set of objects	[32,88–91]	Able to learn a "similarity" measure between the segmentations of two images
Thresholding-based	– Separate objects from background – Fast and small memory needed	Sensitive to nonstationary and correlated noise, ambient illumination, and contrast	[92]	Critically affects the performance of successive steps such as classification and retrieval
Shape-based	– Statistical models – Deformable models	– Handling overlapping objects – Initialization of shape curves – Sensitive to object occlusions – Presence of outliers	[93–101]	Identification of objects of interest

the XYZ color space to achieve perceptual uniformity. As in YCbCr, L^*a^*b consists of one lightness dimension (L) and two chrominance dimensions (a^*, b^*) based on the color opponent process.

However, color histograms have no information about the spatial distribution of color; therefore, other representations have been proposed such as color correlogram and auto-correlogram. These methods provide information about how the spatial correlation of color pairs changes with the distance in an image, and they have shown better retrieval performance than color histograms [105]. In addition, many algorithms have been proposed and used in the CBIR domain based on using color moments. The key idea of color moments is the use of standard deviation and mean of distributions in each color band as a color feature. This considered as a compact characteristic therefore it is usually used as an optimization process along with other color features. Pseudo-Zernike moments [108] have good properties of orthogonally and rotation invariance. Furthermore, it has been validated that Pseudo Zernike moments are superior to Zernike moments in terms of feature representation [109]. Xiaoyin [104] has proposed a new color image retrieval method using the color moment invariant. Representative colors are computed for each image instead of being fixed in a given color space, thus this allows feature representation to be more accurate and compact. The proposed method is based on small image descriptors and adaptive to the context of the image itself by two-stage clustering approach.

Another method has been developed is the color co-occurrence matrix (CCM) [110] which takes into account the spatial relation between color channels. An image can be considered as a composition of suitable "elementary structures". The elements of those pixels carry visual attributes, i.e. colors, and possess relations, i.e. distances between them. Consequently, image contents can be characterized by an appropriate M-dimensional CCM where the attributes and relationships are represented by different matrix axes. Jhanwar et al. [111] have employed another method in the CBIR domain, namely motif co-occurrence matrix (MCM), which is conceptually similar to the CCM. The image is subdivided into 2×2 pixel grids, and each grid is replaced by a scan motif which minimizes the local gradient while traversing the 2×2 grid and forming a motif transformed image. The MCM matrix is then formulated as a 3D matrix where the entry (i,j,k) indicates the probability of finding a motif (i) at a distance (k) from a motif (j) in the transformed image. Guoping [112] has proposed the block truncation coding (BTC) as a different image coding technique which has been employed in the CBIR for compressing color images. From the BTC compressed stream without decoding, two description features of image content have been derived: block color co-occurrence matrix (BCCM) and block pattern histogram (BPH). Both BCCM and BPH have been used to compute the similarity measures of images for CBIR applications. Recently, some BCT variants [113,114] have been proposed for CBIR tasks.

The dominant color descriptor (DCD) [115] has been widely applied in image retrieval applications as one of MPEG-7 color descriptors, which represents the color information of the whole image by a small number of representative colors. The DCD describes the representative color features and distributions in the image or regions of interest through an intuitive, effective, and compact format. Hong et al. [116] have proposed another method based on the fixed number's MPEG-7 DCD. The feature extraction process does not require a threshold value and uses eight fixed dominant colors. The histogram intersection algorithm is utilized to measure features and simplify the similarity computation complexity. Rui et al. [117] have used the DCD along with the fuzzy support vector machine (FSVM) to solve the common problems encountered by the conventional SVMs: small size of the samples, biased hyper-plane, over-fitting, and real-time limitations. Zeng et al. [118] have recently used the color coherence

vector (CCV) which is based on the distance histogram. They have also proposed a multiscale distance coherence vector (DCV) algorithm in accordance with problems where different shapes have the same descriptor and the poor performance of anti-noise of image retrieval algorithm based on the DCV. This algorithm obtains relatively smoother contour curves by Gaussian function evolving contour curve, and then calculates the DCV of the original contour curve and evolved contour curves. This algorithm is invariable to the translation, rotation, and scaling transformations.

However, image noises can arise from numerous sources during image formation and transmission. If the image is destined for human consumption, the noise will reduce the perceptual quality of the image thus its inherent value will be limited. Similarly, if the image is destined for a numerical analysis, the noise will usually limit the system performance, if does not defeat it altogether [119]. Therefore, noise filtering, which is the process of estimating the original image information from noisy data, could be beneficial as a preprocessing for CBIR systems.

2.2.1.2. Texture. In computer vision, there is no precise definition of image texture, but it can be defined as all what is left after considering colors and shapes, or as a description of image structure, randomness, granulation, linearity, roughness, and homogeneity. Image texture is an important image feature for describing innate surface properties of a particular object and its relationship with the surrounding regions [120]. Since texture characteristics are presented in many real images, they are very important and beneficial in pattern recognition and image retrieval tasks. However, the computation complexity and retrieval accuracy are the main challenging drawbacks of texture-based image retrieval systems.

Many texture-based image retrieval methods have been proposed and improved in the CBIR context. Some commonly used algorithms as texture descriptor are Gabor filters, Wavelet transforms, gray-level co-occurrence matrix (GLCM) [121], Markov random field (MRF) [122], edge histogram descriptor (EHD) [123], steerable pyramid decomposition (SPD) [124], and Tamura features [125]. Gabor filters are set of wavelets and each wavelet captures energy at a specific frequency and orientation. Gabor wavelet transforms have multi-resolution and multi-orientation properties and this is optimal for measuring local spatial frequencies [126]. Expanding a signal using this basis provides a localized frequency description and capture local features/energies of the signal. The scale (frequency) and orientation are useful for texture analysis. Gabor elementary functions are Gaussians modulated by complex sinusoids [127]. Lianping et al. [128] have discussed the effects of using a number of Gabor parameters (i.e. number of scales/orientations and filter mask size) on texture-based image retrieval. Many transform-based feature extraction techniques have been be also applied, including discrete wavelet transform (DWT), discrete cosine transform (DCT), Walsh transform, Fourier transform and 2D moments. The DWT is one of the common transforms applied to image processing and retrieval applications. It is used to extract low-level features due to its superiority in multiresolution analysis and spatial frequency.

Many recent works in the CBIR utilize these algorithms either by using a single descriptor or combining many algorithms to form a robust descriptor for image texture. Xingyuan and Zongyu [129] have proposed a structure element descriptor (SED) to extract and describe the image texture and color. The structure elements are defined by five structure elements indicating five directions, respectively. The structure element histogram (SHE) is computed by the SED, and utilizes the HSV color space that has been quantized to 72 bins. The SEH combines the advantages of both structural and statistical texture description methods, and it is able to represent the spatial correlation of color and texture. Liu et al. [130] have developed a new image retrieval approach, namely micro-structure descriptor (MSD). The micro-structures are defined by the edge orientation similarity with the underlying colors that can effectively represent image features. The underlying colors are colors with a similar edge orientation, which can mimic the human color perception as well. With a bridge of microstructures, the MSD can extract and describe texture, color and shape features simultaneously. The MSD integrates the advantages of both structural and statistical texture description approaches. Moreover, this algorithm simulates the mechanism of human visual perception to a certain extent. The MSD algorithm has a high efficiency and indexing performance for image retrieval, but with lower dimensionality of only 72 for full color images. Another approach has been proposed by Chatzichristofis et al. [131], namely the fuzzy color and texture histogram (FCTH) which is formed by the integration of 3 fuzzy systems. The FCTH size is limited to only 72 bytes per image, thus it is suitable for large-scale image databases. The proposed feature is appropriate for image retrieval even in distortion cases such as noise, deformations and smoothing.

Kwitt et al. [132] have introduced a probabilistic texture retrieval approach. It is based on the image representation in the complex wavelet domain and several statistical models for the magnitude of the complex transform coefficients. Additionally, this approach includes closed-form expressions for the KL-divergences between the proposed statistical models which allow constant complexity similarity measurements. In [133], a framework of texture image retrieval as a new family of stochastic multivariate modeling has been proposed which is based on Gaussian copula and wavelet decompositions. They have utilized the copula paradigm to separate a dependence structure from a marginal behavior, and introduced two multivariate models using the generalized Gaussian and Weibull densities. These models capture both the subband marginal distributions and the correlation between wavelet coefficients. In addition, they have derived, as a similarity measure, a closed form expression of Jeffrey divergence between Gaussian copula-based multivariate models. Wang et al. [134] have presented a texture image retrieval method based on the CCM feature. Their approach obtains the color connectivity region set for a colorful image, and then extracts the co-occurrence matrix for 4 orientations (horizontal 0°, vertical 90° and diagonal 45° and 135°) for each connectivity region. The obtained feature reflects the texture correlation as well as represents the color information. Therefore, this method is considered a superior to the GLCM and color histogram and provides a better retrieval performance for texture images. Lai et al. [135] have presented a user-oriented framework in an interactive CBIR system based on the interactive genetic algorithm (IGA). Color distributions, standard deviation, mean value, and image bitmap are used as an image color descriptor. In addition, the entropy based on the EHD and the GLCM is considered as a texture descriptor for image characterization. In particular, the IGA can be used as a semi-automated exploration approach with user help to navigate and identify a complex universe of images with maximum user satisfaction.

2.2.1.3. Shape. Image shape feature basically carries semantic information and can be broadly categorized as boundary-based and region-based. The boundary-based method extracts features based on the outer boundary of the region while the region-based extracts features based on the entire region [136]. Generally, shape-based retrieval methods suffer from problems associated with the translation, scaling, rotation invariances and the stability with slight changes in shape. In consequent, shape descriptors are usually extracted and used with other features such as color and texture and tend to be efficient in specific applications such as man-made objects [14]. Shape descriptor can be represented using many common methods such as polygonal approximation, Fourier

descriptors, invariant moments, deformable templates, B-splines, curvature scale space (CSS), aspect ratio, circularity, and consecutive boundary segments [15,137].

Liu et al. [138] have proposed a new image feature representation method for image retrieval, namely the multi-texton histogram (MTH). The MTH approach utilizes the benefits of the CCM and histogram by representing the attribute of the CCM using a histogram. As a shape descriptor, the proposed MTH method is mainly based on Julesz's textons theory [139] and more efficient than representative image feature descriptors such as the texton co-occurrence matrix and edge orientation auto-correlogram. Another interesting work has been developed by Bronstein et al. [140], namely the Shape Google. This approach has been proposed in the context of non-rigid shape retrieval, and inspired by the work version of Ovsjanikov et al. [141], where the approach was first introduced. Based on heat kernels of the Laplace–Beltrami operator, they show the feature detector and descriptor that are used to construct the vocabulary of geometric words and distributions which serve as a shape representation. This representation is robust under a wide class of perturbations, invariant to isometric deformations, and allows comparing shapes undergoing different deformations. This strategy considers the spatial relations and represents the shapes as compact binary codes that can be efficiently indexed and compared using the Hamming distance. Xiang-Yang et al. [136] have proposed an image retrieval scheme by combining three features: texture, color and shape information, to achieve higher retrieval efficiency. For the color descriptor, the fast color quantization algorithm with clusters merging is used to predetermine the image, and then it obtains a small number of dominant colors with their percentage. For the texture descriptor, spatial texture features are extracted using the steerable filter decomposition which is a flexible approximation method. For the shape descriptor, pseudo-Zernike moments of the image are used to provide a better feature representation due to its robustness against image noise better than other moment representations.

In general, contour-based shape methods require a high computation time because of obtaining the correspondence between contour points from two shapes respectively using local contour information. To solve this problem, Shu et al. [142] have proposed a new contour-based descriptor for closed curves, namely the contour points distribution histogram (CPDH), which depicts the deformable potential at each point along a curve. In addition, they have developed a ground distance calculation technique, which is based on the earth mover's distance (EMD) under polar coordinates, for shape matching in order to be invariant to scale and translation. Another approach, that combines the contour-based and region-based shape methods, has been proposed by Chen and Xu [143], namely the rolling penetrate descriptor. This method improves the conventional methods by obtaining any desired information in a unified way rather than in a specific aspect of shape features. Since different feature functions represent different shape features, the scanning process either (1) acts as a contour descriptor when the feature function calculates the distance between the boundary point and centroid, or (2) describes the relationship between the moment of inertia along the scan line and the angle θ when the feature function accumulates the product of each point and its square distance to the scan line. The feature function is insensitive to noise, resistant to distortion, and its scanning process remains the same regardless of the shape complexity [143].

2.2.1.4. Spatial information. Most of traditional low-level features described earlier lack of spatial information in their extracted representation, e.g. histograms and shape points. Two different parts in the same image may have the same histograms, but with different spatial distribution. Consequently, using an abstract representation alone is not sufficient to represent the pictorial semantic content of images. Regions of interest (ROIs) and graph/tree based representations gained more attention recently as they largely provide vital spatial information which is especially necessary in region-based image retrieval. Other spatial-based mechanisms have been developed for some CBIR applications such as the use of strings to represent the complex topological relations between objects [144], and the use of matrices to indicate spatial relations and directions between objects [145].

Many methods [146–148] divide the entire image into a set of blocks and so allow for ROIs identification, and they employed different overlapping and indexing mechanisms by storing the spatial location of each block or ROI as an index. However, these methods employ a fixed size of regions and do not consider multiple ROIs for the similarity matching between different ROIs of different images. Multiple ROIs [149–151] are employed to provide relative locations of multiple ROIs, and it considers the other blocks that have different spatial locations from ROIs in the image. Lee and Nang [149] have used the MPEG color dominant as feature extracted from image blocks, and selected the blocks having a higher overlapping area to overlap them with user's identified regions. The similarity weighting is based on the relative locations between the query image and target images. To provide a further detailed level of relative location similarity, Shrivastava and Tyagi [120] have incorporated a retrieval method based on region codes for different regions in the image. Region codes along with dominant color and texture features are combined and indexed. Region codes are used for similarity comparison and further used to find relative locations of multiple ROIs in the query and target images.

Graph-based spatial representations are also widely utilized in many image retrieval and recognition applications. Graphs can be efficiently utilized for the similarity of spatial arrangements, where the individual objects or regions are represented by graph nodes and their relationships are represented by arcs between nodes [152]. Alajlan et al. [153] have developed a curvature tree-based (CT) framework for geometry-based image retrieval. It includes the shape and topology of objects and holes composing an image, and the similarity between multiobject images is measured based on the maximum similarity subtree isomorphism between their CTs. Hoang et al. [154] have introduced an image content representation describing the spatial layout with triangular relationships of visual entities. Bunke and Riesen [155] embedded a given graph population in a vector space to interpret the distances of the graph to a number of prototype graphs as numerical features. Kumar et al. [156] have developed a graph-based framework applied in medical CBIR which represents the relationships of multi-modality image contents on a complete graph. The similarity between query and database images is computed upon the spatial locations of image contents. However, graph-based spatial representations are computationally expensive. Table 2 summarizes the global image features along with their main characteristics, limitations, and semantic contributions.

2.2.2. Local image features

Local image descriptors describe local information using key points of some image parts such as region, object of interest, edges, or corners. Recently, local descriptors have shown their superiority for various types of applications in computer vision, and they also have advanced the research efforts in CBIR domain. Local extracted descriptors have many advantages over traditional global features since they are invariant to image scale and rotation, and provide robust matching across a wide range of different situations [157]. As the research attention in CBIR context has shifted to use these local features, we will present and discuss the most common and recent algorithms along with their variants.

Table 2
The main characteristics of global image features.

Features	Main attributes	Limitations	Examples	Semantic contributions
Color	Independent of image state (e.g. size and direction)	– Limited spatial information – Lack of perceptual similarities	– Histograms – Moments – Correlograms – Co-occurrence matrix	It could be employed for image labeling and retrieval based on high-level color semantics
Texture	Description of image structure, randomness, granulation, linearity, roughness, and homogeneity	– Noise sensitivity – Computation complexity	– Gabor filters – Wavelets – GLCM, MRF – EHD, SPD – Tamura features	Describes innate surface properties of an object and its relationship with the surrounding regions
Shape	Binary representation of image objects	Sensitive to translation, scaling, rotation invariances and stability	– Fourier descriptors – Polygonal approximations – Invariants moments – Deformable templates – B-splines, CSS	Carries semantic information based on boundaries and regions
Spatial Information	Spatial arrangements that describe relationships within images	– Computation complexity	– ROIs – Geometric-based – Graph-based – Topology-based	Local and spatial configurations of objects are considered to mimic how humans perceive and compare multiobject images

2.2.2.1. Scale-invariant feature transform (SIFT). One of the most successful and widely applied local image descriptors in the last decade is the SIFT descriptor, which was first introduced by David Lowe [157]. The SIFT includes both key points detector and descriptor that is very robust for extracting distinctive invariant features from images in order to perform reliable matching between different views of objects or scenes. These features are invariant to image scale and rotation and provide robust matching across a fundamental range of noise addition, affine distortion, and changes in both illumination and 3D viewpoint. To find the position and scale, the SIFT detects the locations of key points in the scale space of the image using the scale space extrema in the difference-of-Gaussian (DoG) function included in the image, i.e. $D(x,y,\sigma)$, which is calculated from the difference of two nearby scales separated by a constant multiplicative factor k as follows:

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \quad (1)$$

To hold the rotation invariance, a reference direction is chosen based on the direction and magnitude of the image gradient around each point. A vector which contains values of all orientation histogram entries forms the descriptor, see Fig. 3. The standard SIFT uses a 4×4 array of histograms with 8 orientation bins in each, i.e. $(4 \times 4) \times 8 = 128$ element feature vector for each key point. This feature vector is then normalized to a unit length in order to reduce the effects of illumination changes.

Ke and Sukthankar [158] have proposed a variant of SIFT, namely the PCA-SIFT. Like SIFT, the descriptors in this improved method encode the salient aspects of the image gradient in the feature point's neighborhood. However, instead of using weighted histograms as in SIFT, the principal components analysis (PCA) is applied to the normalized gradient patch to reduce the dimension. Since the PCA is well-suited for representing key point patches, the PCA-SIFT descriptors are more compact than the standard SIFT representation, more distinctive, and more robust to image deformations. Consequently, using this alternative representation of SIFT in image retrieval applications provides faster matching and increased accuracy. Another extension to SIFT has been proposed by Mikolajczyk and Schmid [159], namely the gradient location-orientation histogram (GLOH). It computes the SIFT descriptor for a log-polar location grid with 8 in angular direction, which forms 17 location bins, and three bins in radial direction, but the central bin is not divided in angular direction. The gradient orientations are quantized into 16 bins which forms a histogram of 272 bins. The PCA reduces the descriptor size and the largest 128 eigenvectors are used for description. In [159], a comparison has been introduced on the performance of some descriptors (e.g. steerable filters, shape context, SIFT, PCA-SIFT, spin images, differential invariants, complex filters, cross-correlation, and moment invariants) computed for local interest regions of images in terms of affine transformations, scale changes, rotation, blur, JPEG compression, and illumination changes. As stated, experimental results have shown that SIFT-based descriptors outperform others.

2.2.2.2. Speeded Up Robust Features (SURF). It is another robust local feature detector which was first introduced by Bay et al. [160]. Generally, the high dimensionality of SIFT descriptor is a drawback

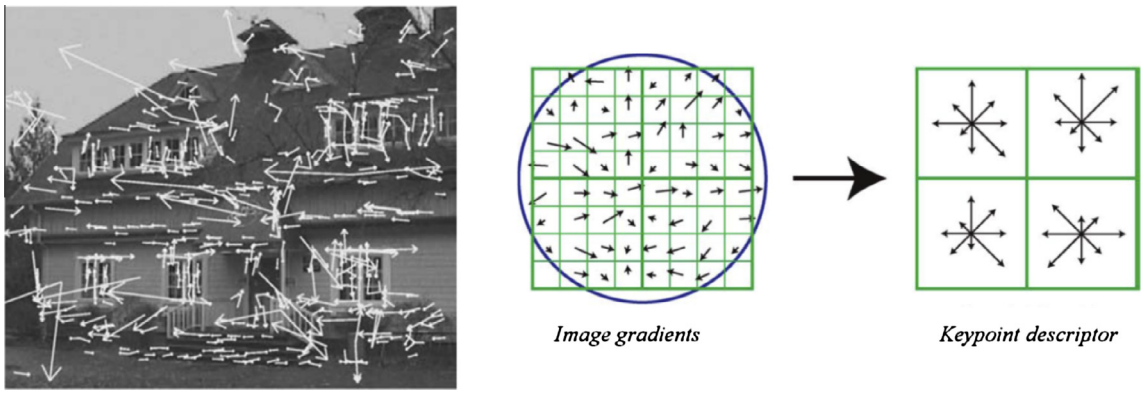

Fig. 3. Sample of SIFT detector and descriptor [157].

at the matching step, and applying PCA in PCA-SIFT [158] and GLOH [159] slows down the feature computation. This descriptor is partly inspired by the SIFT descriptor so the authors have claimed that the standard SURF is faster than SIFT and more robust against different image transformations. Basically, the SURF is mainly based on sums of 2D Haar wavelet responses and makes an efficient use of integral images. The detector of interest points in the SURF is based on the Hessian matrix which detects blob-like structures at the locations where the determinant is the maximum. Unlike the Hessian–Laplace detector, this method relies on the determinant of the Hessian for the scale selection. Additionally, the SURF descriptor describes the distribution of intensity content within the interest point neighborhood, which is similar to the gradient information extracted by SIFT and its variants. It is built on the distribution of first order Haar wavelet responses in x and y direction rather than the gradient, and exploits integral images for more speed by using only 64-D feature size. The SURF also includes a new indexing step based on the Laplacians sign which reduces the time needed for feature computation and matching, and increases the robustness simultaneously.

2.2.2.3. Local patterns. Some other methods based on local patterns of qualitative level differences have been proposed. The local binary patterns (LBPs) have been presented by Ojala [161] as one of the widely used approaches for texture discrimination. It derives a generalized grayscale and rotation invariant operator presentation that detects the "uniform" patterns for any quantization of angular space and for any spatial resolution. Additionally, it presents a method for combining multiple operators for multiresolution analysis. The LBP approach uses eight neighboring pixels and the value of the center pixel as a threshold, and these values form the LBP code after some weighted calculations. Since the operator is invariant against any monotonic transformation of the gray scale, the LBP is very robust. In addition, the operator can be realized with a few operations in a small neighborhood and a lookup table, this result in the advantage of computational simplicity. Recently, some other LBP variants have been proposed in the literature. Liao et al. [162] have extended the LBP method to the dominant-LBP (DLBP). Since the uniform LBPs are not the dominating patterns in some textures with irregular shapes and edges, this motivated them to propose the DLBP feature extraction method which is robust to histogram equalization and rotation. Unlike the LBP which only employs uniform LBPs, the DLBP approach obtains the occurrence frequencies of all rotation invariant patterns defined in the LBP groups. After sorting these patterns in a descending order, the first several most frequently occurring patterns are the dominant patterns in that image. Accordingly, using the DLBP approach is more reliable to represent the dominating pattern in texture images.

Since the LBP has a limitation of losing the global spatial information, Guo et al. [163] have proposed an efficient global matching scheme which uses the LBP for feature extraction, namely the LBP variance (LBPV). This approach does not extract the locally rotation invariant LBP as in [161,162], but instead it builds a rotation variant LBP histogram and then performs a global matching procedure using an exhaustive search. These schemes find the minimal distance in all candidate orientations, even it is computationally extensive. The principal orientations can be estimated by the extracted LBP features and thus matching distances are computed only along these orientations. Unlike the LBP, which computes the joint histogram of LBP and rotation invariant variance (VAR) globally, the LBPV computes the VAR from the local region and accumulates it into the LBP bin. Consequently, the LBPV operator could greatly reduce the requirement for a large number of training samples. Gue et al. [164] have also developed the completed LBP (CLBP) scheme. The CLBP represents the local region by local difference sign-magnitude transforms (LDSMT) and its center pixel. The image gray level is represented by the center pixels that are converted into a binary code, referred as CLBP-center (CLBP_C), by global thresholding. The CLBP_C is combined with two operators, namely CLBP-sign (CLBP_S) and CLBP-magnitude (CLBP_M), into joint or hybrid distributions, and a significant improvement can be achieved for rotation invariant texture classification.

Since the LBP is considered as nondirectional first-order local patterns collected from first-order derivatives, Zhang et al. [165] have proposed the local derivative patterns (LDPs) approach, and extended the LBP to the nth order LDPs. However, LBP,LDP, and their variants are sensitive to appearance variations (e.g. illumination, pose, and facial expression) that usually occur in unconstrained natural images. In order to address this problem, Tan et al. [166] have introduced the local ternary pattern (LTP) which has been employed for the face recognition on different lighting conditions. However, LBP, LDP, and LTP extract the information based upon the distribution of edges which are coded by only two directions (i.e. positive or negative). To improve the performance of these methods, another successful algorithm which is based on four direction codes, namely the local tetra patterns (LTrPs), has been proposed by Murala et al. [167] which effectively used in the CBIR domain. Recently, Jeena et al. [168] have proposed the local oppugnant color texture pattern (LOCTP) which is inspired by the LTrPs to improve the retrieval performance. The LOCTP operator is obtained by computing the texture pattern over three channels of the oppugnant color space. It could be considered as a joint color-texture descriptor to extract uniform and non-uniform color texture features. The LOCTP determines the relationship between the referenced pixels and their oppugnant neighbors in terms of intensity and directional information.

2.2.2.4. Histograms of Oriented Gradients (HOG). Generally, some of aforementioned local detectors are computed on the dense grid of uniformly spaced cells, and the performance is improved using overlapping local contrast normalizations. Dalal and Friggs [169] have proposed locally normalized HOG descriptors to provide better performance compared to other existing features including wavelets. Basically, this method characterizes the local object appearance and shape by edge directions or the distribution of local intensity gradients, even without an accurate knowledge of edge positions or the corresponding gradient. The HOG divides the image window into small spatial cells as regions, and accumulates edge orientations over the pixels of each cell or a local l-D histogram of gradient directions. As a result, image representation is formed by the combined histogram entries. Furthermore, it accumulates an "energy" as a measure of local histogram over "blocks" which somewhat larger spatial regions and then the block cells normalized using these results. This provides a better invariance to shadowing, illumination, etc. The human detection chain is obtained by tiling the detection window with a dense grid of HOG descriptors along with the combined feature vector in a conventional SVM based window classifier. In the last decade, the HOG has been used in many applications and proved as a successful method especially for object recognition. Chandrasekhar et al. [170] have proposed another alternative descriptor, namely the compressed histogram of gradients (CHoG), as a low bit-rate descriptor with a 20× reduction in the bit rate. The CHoG represents gradient histograms as compressed tree structures which avoids the need for decoding to compute the distances between descriptors in their compressed representation. This approach has a low complexity and speedup the matching stage.

2.2.2.5. Others. The GIST feature [171] has received increasing attention in the context of image retrieval. It has been introduced as a low-dimensional representation of the structure in real world

scenes, which is called the spatial envelope. It does not require any segmentation or processing on the individual objects or regions, so it extracts the orientation histograms of square image grids. The GIST structure is described as a set of perceptual properties (naturalness, openness, roughness, ruggedness and expansion). These properties are related to the shape of the space and meaningful to human observers, which provide a holistic description of the scene where the local object information is not taken into account. The GIST organizes scene pictures as the human subjects do, and it is able to retrieve images that share the same semantic category. Therefore, it provides a meaningful representation of complex environmental scenes that may sketch a direct interface between the visual perception and the semantic knowledge.

The difference between pixels of scan pattern (DBPSP) is another feature used in the CBIR domain, which mainly calculates the differences among all pixels within motifs of a scan pattern. Specifically, it records pixel value differences among all scan directions within motifs of scan pattern, and then takes the appearance rate of total pixel value differences in the whole image as a feature. The DBPSP is usually combined with other image features and used as texture feature such as in [172]. As an alternative to SIFT and SURF, Rublee et al. [173] have introduced a binary descriptor called the oriented fast and rotated brief (ORB), which is based on FAST detector [174] and BRIEF descriptor [175]. The ORB adds an orientation component to FAST, and learning method for de-correlating BRIEF features under a rotational invariance.

Sliding windows is another approach used to extract local image objects. It evaluates a quality function (e.g. a classifier score) over many rectangular sub-regions (i.e. windows) of image and taking its maximum as object location. Object localization, where the target is to find a bounding box around the object, provides a ground truth annotation for bounding boxes better than pixel-wise segmentations. However, sliding windows based methods criticized as computationally expensive due to slow search mechanisms employed and high memory usage. To optimize the localization, Lampert and Blaschko [176] have proposed the efficient subwindow search (ESS) algorithm to find the globally maximal region, independent of the shape of quality function. The ESS is robust and performs faster on object localization and localized retrieval, and allows the utilization of local classifiers.

Table 3 summarizes the main attributes of local image features. However, CBIR systems need to form and index these local image descriptors in order to match them with visual data in the database. For example, the key point descriptors obtained using SIFT, which are vectors with variable sizes, could be formed as image feature using the bag-of-words (BOW) approach. In the next subsection, we will discuss image representation and indexing of image features extracted; especially in high-dimensional space which need to be reduced.

2.3. Dimensionality reduction and indexing

Despite the considerable volume of algorithms conducted for dimensionality reduction, the 'curse of dimensionality' still challenges the field of CBIR. To form more discriminating image descriptors, further research efforts on indexing algorithms and structures are expected in order to effectively associate low-level features with further semantic information. However, describing images in higher semantics usually leads to generate high-dimensional image features which predominantly have a sparse distribution of data. This obviously will degrade the retrieval performance of CBIR systems. Consequently, the dimensionality reduction is a necessary solution for this problem. The goal here is to project data into another space to highlight certain structures by identifying only the interested projections. Many methods have been proposed in the literature and discussed by some

Table 3
The main characteristics of local image features.

Features	Main attributes	Limitations	Examples/Variants	Semantic contributions
SIFT	– Invariant to scale and rotation – Robust matching across a range of noise addition, affine distortion, and change in illumination/ viewpoints	– High-dimensional at matching – Need to be encoded in fixed-size vectors for image matching	PCA-SIFT [158] GLOH [159]	Preserves only interest points that are likely to remain stable over transformations, and builds distinctive descriptions
SURF	– Hessian matrix-based detectors – Relies on integral images to reduce the computation cost	– Poor approximation of key-point orientations – Poor performance on a rotational invariance	Accelerated SURF [177]	Same as SIFT
Local Patterns	– Robustness to monotonic grayscale changes – Computational simplicity	Sensitivity to noise in near-uniform image regions	DLBP [26], LBPV [163], CLBP [164], LDP [165], LTP [166], LTrPs [167], LOCTP [168]	Extracts uniform and non-uniform discriminative texture features
HOG	– No need of accurate edge positions – Local contrast normalization	Multiple bounding boxes at object detection	CHoG [170]	Characterizes the local object appearance and shape
GIST	– Low dimensional descriptor – No need for preprocessing	Low degree of invariance	Compact GIST [178]	Provides perceptual properties and meaningful to human (e.g. naturalness, openness, roughness, ruggedness and expansion)
DBPSP	Describes the direction and complexity of textures	Sensitive to noise, size, and rotation variants	Texture [172]	Describes the texture distributions
ORB	– Binary-valued and compact feature – -Robust to lighting, blur, and perspective distortion	Low degree of scale invariance	Approximate matching [179]	Same as SIFT and SURF
Sliding windows	– Simplicity of parameterization – Region-based localization	– Computationally expensive – Sensitive to quick variations of objects	EES [176]	Efficient in subsequent scene understanding

comprehensive studies [19,180–182]. In the following, we discuss the popular methods of dimensionality reduction and indexing in the field of CBIR with some important advances on these tasks.

2.3.1. Principal component analysis (PCA)

It is a successful and widely used method for dimensionality reduction. The PCA is a linear transformation method which projects input vectors into new ones using an orthogonal matrix (i.e. eigenvector of sample covariance matrix). Based on estimated corresponding eigenvector, the vector components are then calculated as orthogonal transformations called principal components. The number of principal components is reduced by considering only the first several eigenvectors that sorted in a descending order of the eigenvalues, thus it mainly relies on Gaussian features. As a generalized form of PCA, the Kernel-PCA (KPCA) has been introduced by Scholkopf et al. [183]. The KPCA is a nonlinear method and based on the kernel method to produce (K) as a kernel matrix. It maps input vectors into a high-dimensional feature space and then computes the linear PCA, so it is simple for nonlinear data processing.

Like the PCA, the independent component analysis (ICA) and singular value decomposition (SVD) are also other linear approaches for dimensionality reduction. The ICA exploits inherently non-Gaussian features and employs higher moments of measured data, thus it is preferred when data cannot be ensembled and very noisy. Unlike the PCA, which minimizes the covariance of the data, the ICA minimizes the mutual information of the output, i.e. identifies the independent components for non-Gaussian signals. The SVD is also widely used in determining the principal components of a multi-dimensional data. Consider a real $m \times n$ matrix X of observations which may be decomposed as follows:

$$X = USV^T, \tag{2}$$

where S is a non-square matrix with zero entries everywhere, except on the leading diagonal with elements S_i arranged in descending order of magnitude. Each S_i is equal to $\sqrt{i}$, i.e. the square root of eigenvalues of matrix $C = X^TX$. A stem-plot of these values against their index i called the singular spectrum. The smaller eigenvalues have lower energy along the corresponding eigenvector. Therefore, the smallest eigenvalues are often considered to be due to noise. The matrix V is a $n \times n$ matrix of column vectors which are the eigenvectors of C. The $m \times m$ matrix U is the matrix of projections of X onto the eigenvectors of X. If the most significant k eigenvectors are only retained, then the truncated SVD of X is given by $Y = US_kV^T$.

2.3.2. Bag-of-word (BOW)

It is one of the most popular feature representations in the CBIR context. This method is extended from text-based retrieval systems. The BOW [184] assigns the image descriptors extracted (e.g. SIFT) to the closest visual words in a visual vocabulary, i.e. 'codebook'. The k centers of codebook's clusters are computed and learned by the k-means clustering method. This high-dimensional sparse vector represents the image and weighted using the term frequency inverse document frequency (tf-idf). The similarity between BOW vectors can be computed by the standard similarity distances, such as the Euclidian or Manhattan measures. Despite that the BOW image vectors provide better semantic representation than low-dimensional vectors; it needs an efficient indexing method to reduce the unfavorable effect of high-dimensionality on retrieval performance. Recently, Jegou et al. [181] have proposed a moderate dimensionality of image representation to achieve better search performance in large-scale image retrieval. They have used Fisher kernels [185] algorithm to aggregate variable-size local image descriptors into a compact vector representation. These vectors are reduced to a few hundred components by applying the PCA projection and compared using the standard L2 distance. Finally, the aggregated vectors are indexed and compactly encoded by the approximate nearest neighbors search method.

2.3.3. Fisher kernel

Most of standard BOW representations suffer from the sparsity and high dimensionality representations. The Fisher kernel provides a compact and dense representation which is necessary for image classification and retrieval applications. Fisher kernel is a probabilistic model that identifies the similarity between objects using sets of measurements for each object with a higher order of statistics than the BOW, which efficiently processes a variable length of data. Due to its lower computations and simplicity, Perronnin and Dance [186] have applied Fisher kernels on image classification and large scale image search. Jegou et al. [187] have introduced a simplified non-probabilistic version of Fishers kernels, the vector of locally aggregated descriptors (VLAD). The VLAD image representation is also trained using k-means to accumulate the local descriptors and then normalized by L2. Belhumeur et al. [188] have presented a supervised linear dimensionality reduction approach, namely the Fisher linear discriminant analysis (FLDA). The FLDA maximize the inter-class scatter matrix and minimize the intra-class scatter matrix; so the eigenvectors matrix of image features extracted is obtained as a result. However, to form the resultant intra-class scatter matrix as a nonsingular, the PCA is used to reduce the dimension of input features. Rahulamathavan et al. [189] have recently used the local fisher discriminant analysis (LFDA), as extension of FLDA, to perform facial expression recognition in the encrypted domain.

2.3.4. Manifold learning

A nonlinear approaches that are usually used in the learning process and simplify data as embedded manifold (lower dimensionality) within a higher-dimensional space, i.e. the original data visualized in a low-dimensional space. Some representative methods of manifold learning are: manifold sculpting (MS) [190], isometric mapping (ISOMAP) [191], locally linear embedding (LLE) [192], and locality preserving projections (LPP) [193].

2.3.5. Tree-based indexing

Many indexing techniques represent the data space of images into a tree hierarchical structure which can be categorized as tree-based methods [19]. The non-leaf nodes are directory nodes where the information of data space is stored, and the leaf nodes are data objects that store the information to be indexed. To efficiently index high-dimensional descriptors by tree-based indexing methods, the conventional reduction methods are used as a preprocessing step. Some recent representative works on tree-based indexing are KD-tree [194], R-tree [195], M-tree [196], and EHD-tree [197].

2.3.6. Hash-based indexing

To enable fast processing of equality selection queries, an access method is required to group the original data by their value on some attributes. The hash-based scheme maps the search-key values with a collection of buckets so that their mapped values are determined by a function called the hashing function [198]. The hash-based high-dimensional indexing projects data features from high dimensions to low dimensions via those hash functions that can be broadly categorized into two groups: sensitive hashing and spectral hashing. Sensitive hashing methods such as the local sensitivity hashing (LSH) [199] project feature points into a low-dimensional Euclidian space, while spectral hashing methods [200] map the close data points in the Euclidian space to similar

binary codes in a low dimensional Hamming space [19]. Some of recent works on hash-based indexing are multi-probe KLSH [201], semantic hashing [202], sparse spectral hashing (SSH) [203], and spherical hashing [204].

2.3.7. Latent semantic indexing (LSI)

The LSI has been developed by Deerwester et al. [205]. It has been originally used as a mathematical technique for text retrieval where two documents may semantically close even if they do not share particular keywords. A comparable problem holds in the context of image retrieval as various visual keywords/features can represent the same object but have different meanings. For image retrieval, the matrix is constructed in the similar concept except that "documents" are images and "terms" are image features. The LSI mainly relies on the SVD algorithm to identify patterns/relationships between the terms and concepts of documents. The LSI has been used by many studies [206,207] to exploit the underlying semantic structure of web images and videos based on their visual features in CBIR domain. However, the major challenges of applying the LSI approach for high-dimensional data, e.g. Web contents, are the high computation time and memory storage required for SVD computations. Table 4 presents the main characteristics of dimensionality reduction and indexing methods.

2.4. Machine learning

CBIR algorithms have been shifted recently to perform the similarity based retrieval on image collections by discovering the underlying structure of images and their descriptors. Such these algorithms do not know in prior into which semantic group the image data fall, these clustering methods are called unsupervised algorithms (e.g. *k*-means clustering). In contrast, if the algorithm knows in advance the image groups and their different semantics, it becomes a classification task in which the algorithm assigns an input image into predefined semantic groups. Generally, classification methods are supervised algorithms (e.g. support vector machines (SVM)). As illustrated in the previous sections, current retrieval methods tend to represent the image by certain local descriptors (e.g. SIFT and LBP). This task is usually followed by classifying image descriptors into some predefined semantic groups by the discriminant analysis, or clustering them by identifying the best boundary that separate descriptors into semantic clusters; i.e. maximizing the similarity within clusters and minimizing the similarity between different clusters.

It is very important to validate the results of clustering and classification, and typically these algorithms perform training and testing processes on some independent data subsets. As a result, the algorithm will learn the underlying properties of data input and generalize the process in order to deal with any new added image descriptors rather than following an explicitly predefined instructions. These concepts highlight the importance of using machine learning strategies along with pattern recognition in the CBIR domain which provides more improved image search. However, this task is challenging both in terms of accuracy and computational cost. Accordingly, many methods and representations have been proposed in the CBIR context with the aim at modeling the learning process.

Since all of the aforementioned concepts are strongly related, here we will review the most widely used algorithms for clustering, classification, machine learning employed in CBIR. In addition, we will discuss some recent promising works that have been proposed, e.g. deep learning and automatic image annotation which will decidedly ameliorate the retrieval accuracy and performance.

Table 4
The main characteristics of dimensionality reduction/indexing methods.

Methods	Attributes	Limitations	Variants/Works	Semantic contributions
PCA, ICA, SVD	– Statistical transformation – Orthogonal vectors from eigenvectors	– PCA is sensitive to relative scaling of original data – ICA cannot determine the actual number/scaling of source signals – Computation cost	Kernel-PCA [183] Fast-ICA [208] K-SVD [209]	– PCA gives a likelihood of data based on the amount of variance – ICA used in blind source separation – SVD transforms feature-image matrices to a "semantic" space of low dimensionality
BOW	Sparse and high-dimensional vectors	– High training cost – Lack of spatial information – Information loss in vector quantization	BOW [184]	Provides a robust semantic representation of image contents
Fisher Kernel	– Simple with low computation cost – Compact and dense representation	Training cost of image representation	VLAD [187] FLDA [188] LFDA [189]	– Provides a 'natural' similarity measure that considers the underlying probability distribution – Efficient on multiclass classification
Manifold learning	– Nonlinear dimensionality reduction – Learn the embedded structure	High computation and learning cost	MS [190] ISOMAP [191] LLE [192] LPP [193]	– Uncovers the intrinsic dimensionality – Enhanced structure description and visualization of image objects
Tree-based indexing	– Flexibility for different shapes after partitioning data space from coarse to fine	Inefficient on high-dimensional space, so depends on dimensionality reduction methods	KD-tree [194] R-tree [195] M-tree [196] EHD-tree [197]	Supports equality selections and range-searches efficiently
Hash-based indexing	Flexible for both data-independent and data-dependent hashing (coding)	Inefficient on sparse descriptors	MP-KLSH [201] Semantic-H [202] SSH [203] Spherical-H [204]	Support equality selections

2.4.1. Clustering and unsupervised learning

Following the feature extraction and representation processes, clustering methods aim to group image descriptors into mutually exclusive clusters with different semantics. The most commonly applied methods are:

2.4.1.1. k-means clustering. By far, it is the most popular clustering method has been used in scientific and industrial applications [25]. Fig. 4 demonstrates the main idea of *k*-means. The key steps in this unsupervised method are: (1) select some initial points from the input data as initial 'means' or 'centroid' of clusters, (2) associate every data point in the space with the nearest centroid to form *k* clusters, (3) recalculate each of initial means and set the new results as 'centroids' of *k* clusters, and finally (4) steps 2 and 3 are repeated until all points of the input data become assigned to a certain cluster.

2.4.1.2. Soft clustering. Despite its simplicity and speed, the *k*-means algorithm generally confronts some challenging issues such as centroids initialization, sensitivity to outliers, and the assignment of some data points that are equally close to many clusters. To deal with overlapping clusters, another two popular extensions of *k*-means have been proposed and widely conducted: Gaussian mixture models (GMM) and fuzzy clustering (e.g. fuzzy *c*-means). The GMM is a probabilistic method which uses the expectation maximization (EM) algorithm to mathematically assign image data points into clusters that are represented as a mixture of multivariate normal densities, i.e. parametric distribution. On the other hand, the fuzzy-based clustering can associate each data point to more than one cluster with a related membership degree between 0 and 1. Both of these methods are usually known as soft clustering [210].

2.4.1.3. Semi-supervised clustering. Other clustering methods employ only a small amount of labeled data which known as semi-supervised clustering. Generally, semi-supervised methods improve the clustering process either by modifying the clustering objective algorithm to satisfy labels or constraints (i.e. constraint-based approach), or by training the distance metric to satisfy the labels or constraints in the supervised data (i.e. metric-based approach). Bilenko et al. [211] have proposed a balanced semi-supervised clustering, the metric pairwise constrained *k*-means (MPCK-means), by unifying the two mentioned approaches. The MPCK-means performs the distance-metric training at every clustering iteration and learns individual metrics for each cluster thus allowing clusters of different shapes. Recently, Papagiannopoulou et al. [212] have introduced a new technique for image clustering by combining a concept-based approach of image representation with clustering techniques. This method uses trained concept detectors to represent each image by a vector of concept detector responses which is then used as input to the clustering algorithms. More specifically, they apply trained concept detectors to the image dataset and receive prediction scores for each concept. As a result, each image can be represented as an element vector of detector confidence scores. After the clustering process, a summary of image collections and events can be formed by selecting one or more images per cluster according to different criteria.

2.4.2. Classification and supervised learning

2.4.2.1. SVM classifiers [213]. It has been widely used as a supervised classifier for image classification and pattern recognition. The SVM decides to which class a new data point will be assigned, and it represents the largest margin between the defined image classes so that the distance between them is maximized. Specifically, the SVM maximizes the margin between the hyper-plane and the nearest data point of each class by indicating the class (Y) to which the data point (x) belongs, where the class is either +1 or −1. The training samples that are close to the hyper-plane called 'support vectors'. Fig. 5 demonstrates the key idea of SVM classifier.

2.4.2.2. Bayesian classifiers. It has been successfully adopted in many computer vision problems including the CBIR domain. For image classification, the general idea is that a set of images is partitioned into classes and any image should belong to one and only one class. Images from class are modeled as samples of a random variable and each class has an a priori probability of membership. The '0/1' loss function specifies the loss incurred when a class is chosen. For image retrieval problems, this decision is based on image features rather than directly on raw pixel values. Therefore, there are independent class-conditional densities for features rather than for raw images, thus the classification problem can be stated as: "given feature sets, classify the image into one of the classes" [214].

2.4.2.3. Ensemble classifiers. The commonly used classifiers including SVM and Bayesian have a large variance, especially if they constructed on small training sets. This degrades the classifier stability and produces a 'weak classifier'. To improve the classifier stability, one of effective solutions is to use a combined decision of many weak classifiers instead of a single classifier. Some of

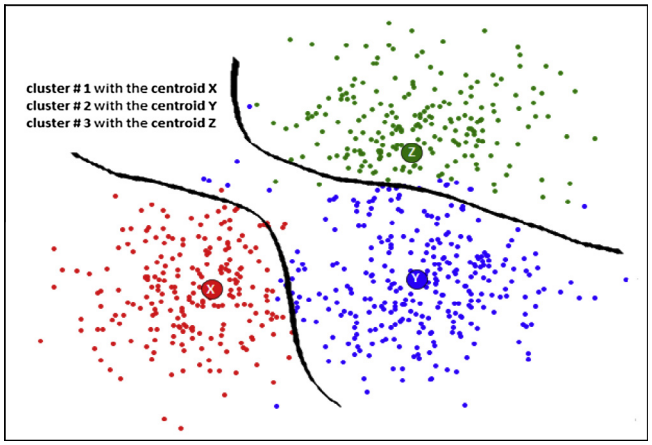

Fig. 4. A simple *k*-means clustering.

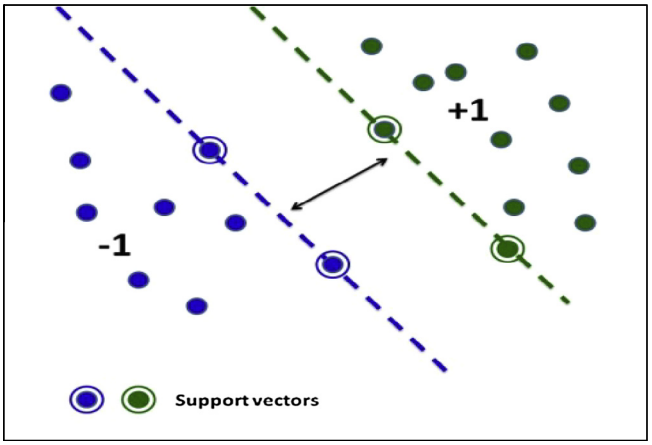

Fig. 5. A standard SVM classifier.

representative methods are: bagging [215], boosting [216], random subspace method (RSM) [217], and rotation forest [218]. These methods are ensemble classification and originally designed for decision trees that constructed by recursively partitioning the input data space into a set of spaces and classifying the features using a set of decision rules. However, these learning methods are generally sensitive to data noise, especially the boosting and rotation forest algorithms. Therefore, Kotsiantis [219] has combined these entire ensemble learning methods (i.e. bagging, boosting, RSM, and rotation forest) with 6 sub-classifiers in each method, and then a voting methodology (e.g. sum rule voting, Vote B&B&R&R) has been used to improve the prediction of the classifier.

2.4.2.4. Artificial Neural Networks (ANNs). Recently, ANNs have been extensively utilized in solving various complex real-world problems including image retrieval. This type of networks has initially developed according to the operational procedure of the human neural system. The attraction of utilizing the ANNs has been raised due to their noteworthy information processing characteristics which are mainly relevant to the high parallelism, nonlinearity, noise and fault tolerance, and learning capabilities [220]. Basically, ANNs compose neurons and connections between them, which together determine the network behavior. Generally, a neuron receives inputs as stimuli from the environment and combines them to form a 'net' input (ξ) which is passed over through a linear threshold gate and transmits the output signal forward to another neuron or environment. The ANNs often consist of three neuron layers: input, hidden and output. The input layer includes n neurons for n pieces of network input signal ($X_1 \dots X_n$) which are independent variables. In the hidden layer, the number of neurons is empirically chosen by the user. Finally, the output layer includes k neurons for k classes which are dependent variables. Generally, each weighted connections between neurons is modified by successive iterations during the network training. In the input layer, the neuron state is determined by the input variable so that other neurons in the hidden and output layers evaluate the signal state of the previous layer as follows:

$$a_j = \sum_{i=1}^{I} X_i W_{ij}, \tag{3}$$

where a_j is the net input of neuron j; X_i is the output value of neuron i in the previous layer; and W_{ji} is the connection weight between the neurons i and j. A specific task can be performed during the learning process in ANNs by updating the internal representation of the system in response to external inputs with a modification in the network architecture. The ANNs learning is iteratively performed as long as the training examples are fed. ANN-based systems are learned to handle noisy, imprecise, fuzzy, and probabilistic information without remarkable counter effect on response quality, and generalize to unknown tasks [199]. The network type selected depends on the problem to be solved, and the backpropagation network is one of the most widely used ANNs. As a classifier, the ANN performs two consecutive stages: training and testing. The extracted features from images are fed into the network as input data and the target data are the class label of images. Therefore, the input and target data are fed into the network to be trained. In the testing stage, a query image is usually used in the same manner to extract its features and form a feature vector which then can be used as input to the trained network which assigns it to the similar classes in the retrieval process.

Park et al. [221] have proposed a method based on neural networks for content-based image classification. They have used object images with foreground and background regions that divided by the JSEG segmentation method. The classification features are shape-based texture descriptors that extracted from wavelet-transformed images. The constructed classifier for image features is based on the back-propagation learning algorithm which depends on the generalized least mean square (LMS) rule to minimize the average difference between the output and the target value in the neural network. Ghiassi and Burnley [222] have developed a dynamic artificial neural network (DAN2) as an alternative to traditional classification methods such as SVM and Bayesian. The DAN2 is based on (1) learning and knowledge accumulation at each layer, (2) adjusting and propagating this knowledge forward to the next layer, and (3) repeating these steps until the desired criteria for network performance are reached. To make this model more sensitive to input variations, they magnified the class values to +100 and −100 as the larger class values aggrandize small differences in the output. Therefore, each point is classified either into positive (the class value of +100) or negative (the class value of −100).

However, the performance of neural networks largely influenced by the architecture determined manually by the trial and error, but it is difficult to determine the best network architecture. Yoon et al. [223] have stated that a knowledge-based ANN usually operates on a prior knowledge from domain experience, which provides better starting points for the target function and achieves a better classification accuracy. Since the identification of prior knowledge is usually difficult, they have developed a new neural network approach, namely the algorithm learning based neural network (ALBNN). The ALBNN improves the classification accuracy by integrating classification procedures with feature selection. Specifically, it employs a prior knowledge instead of using unknown background resources. It also employs the extreme learning machine to obtain better initial points faster and avoid determining architecture and manual tuning which are irrelevant time-consuming works. The ALBNN has shown its ability in producing new relevant features as well as improving the classification accuracy.

However, the existence of noise in the data input that fed into ANNs usually decreases the discrimination and increases both the uncertainty and training error. Therefore, Wu et al. [224] have proposed a vectorization–optimization-method-based type-2 fuzzy neural network (VOM2FNN) to deal with the uncertainty and discriminability. It classifies the noisy data while preserving a small network size. To model and minimize uncertainty effects, the interval type-2 fuzzy sets have adopted in the antecedent parts of the VOM2FNN. To increase the discriminability and reduce the parameters, the vectorization–optimization method (VOM) has been used to tune the consequent parts of VOM2FNN. With extensive research on ANNs in order to improve the CBIR accuracy and performance, the last few years have witnessed a remarkable concentration on using the convolutional neural networks (CNN). Since it has a deep structure, the CNN has been proved as a promising contribution for diminishing the semantic gap. Recently, CNNs have provided considerable achievements in real-word CBIR tasks and approaches such as deep learning which will be discussed in the next section.

2.4.2.5. Logistic regression. Regression-based methods become important in any data analysis concerned with investigating the relationship between the response (i.e. outcome) variable and one/many explanatory (i.e. regressors, predictors or covariates) variables. Linear regression is the simplest model where the outcome variable is assumed to be continuous. This model predicts a target value y starting from a vector of input values $x \in R^n$, and the goal is to find a function such as $y = h(x)$ so that $y(i) \approx h(x(i))$ for each training example. This function is the cost function (i.e. loss or penalty) which measures how many errors incurred while predicting $y(i)$ for a particular choice of θ.

However, this is not the optimal solution for predicting binary-valued labels. Logistic regression is another simple classification algorithm for learning to make such decisions. It predicts the probability that a given example belongs to the first class versus the probability that it belongs to the second one. Additionally, softmax regression is utilized as a generalization of logistic regression to handle the classification problem with multiple classes. Usually, the ordinary least squares (OLS) method finds an unbiased linear combination of each explanatory variable that minimizes the residual sum of squares. However, in case of that regression coefficients being highly correlated or far exceed the sample size, the OLS may yield estimates with a large variance which consequently reduces the prediction accuracy. To solve this problem, regularization-based models such as ridge regression (RR) [225], least absolute shrinkage, and selection operator (LASSO) [226] are widely applied. The RR shrinks the coefficients of correlated predictors equally towards zero, while the LASSO expects many coefficients to be close to zero and only a small subset to be larger and nonzero. The main difference between the classification and the regression is their prediction variables. In classification, the prediction variables are categorical and unordered, while the regression has a numerical or ordered whole values.

2.4.3. Deep learning

Over the last few years, the attempts to mimic the human brain have witnessed some important advances by using a technique known as "deep learning". Deep learning algorithms model the high-level semantics in data by utilizing deep structures which include multiple feature representations and non-linear transformations. Since the human brain is organized in deep and complex architectures and processes data through multiple stages, deep learning methods attempt to stimulate such architectures. Moreover, exploring deep architectures, which automatically learn features from data at multiple levels of abstracts, provides an ability to learn complex tasks that directly map input data to the output without relying on human-made features using the domain knowledge [20]. Therefore, deep learning algorithms have been used recently to reduce the semantic gap and solve many real-world problems in the field of computer vision.

The successful deep learning techniques developed so far share two key attributes [227]: (1) the generative nature of learning model, which typically requires adding an additional top layer to perform discriminative tasks and (2) the unsupervised pre-training step that makes an effective use of large amounts of unlabeled training data for extracting structures and regularities in the input features. However, the deep learning technique composes a wide range of machine learning algorithms and architectures, many layers of non-linear transformations and stages, and different types of neural networks. Therefore, selecting appropriate network architectures depends on how these architectures will be used in the problem domain. Since employing CNNs in the field of CBIR seen as a breakthrough, we will only focus on and discuss deep learning achievements and advances in the CBIR domain that are based on CNNs networks with different tasks, e.g. feature representations learning, distance metric learning, and image annotation.

Krizhevsky et al. [228] have trained one of the largest CNNs (i.e. 60 million parameters and 650,000 neurons) with ReLUs units on a large image dataset which contains 1.2 million high-resolution images that fall into 1000 different classes. As a classifier, their CNN compromises a number of new features which reduce the required training time and improve its performance. The final network contains eight weighted layers: five convolutional and three fully connected. The output of the last fully-connected layer is fed to a 1000-way softmax layer which generates a distribution over the 1000 class labels. To reduce overfitting while learning many parameters in the fully-connected layers, they have performed data augmentation methods along with a regularization method called the dropout. For each hidden neuron, the dropout method sets to zero the output with probability 0.5, thus the neurons, that are dropped out, will not contribute to the forward pass and will not participate in backpropagation. Consequently, the complex co-adaptations of neurons are reduced and enforced to learn more robust features that are useful in conjunction with many various random subsets of other neurons. However, the learning performance in this supervised CNN degrades if the number of layers changed.

Wan et al. [20] have adopted the similar deep architecture of CNNs as proposed in [228] to generalize the trained deep networks for image classification and apply them using the BOW feature representations on certain CBIR tasks. In addition, they have addressed the utilization of a trained CNN-based model in learning feature representation for other CBIR tasks in a new domain, where a small set of training data is available. More specifically, they have applied the trained CNNs to direct feature representation, and taken the activations in the last three fully connected layers (namely FC1, FC2, and FC3) as feature representations. The feature vectors of these direct feature generalization are denoted as "DF.FC1", "DF.FC2", and "DF.FC3", respectively. The feature DF.FC3 is taken from the final output layer, the feature DF.FC2 is taken from the final hidden layer, and the DF.FC1 is the activations of the layer prior to DF.FC2. Instead of directly using the features extracted by the pre-trained deep model, their proposed framework adopts an online similarity learning algorithm, namely the online algorithm for scalable image similarity (OASIS) for similarity learning, which learns a bilinear similarity measure over a sparse representation to minimize the overall loss. Finally, they have retrained the deep models with the similarity learning objective or classification on different domains, e.g. object, landmark, and facial image retrieval tasks. Despite the encouraging retrieval results and performance, the proposed deep framework does not provide a fully generalization and stabilized accuracy over all datasets for different CBIR tasks.

In order to improve the generalization ability, Sun et al. [229] have proposed another deep CNN architecture, referred as the deep hidden identity features (DeepID). It learns a set of high-level feature representations through a deep learning for face verification. The compact and discriminative DeepID has learned through challenging multi-class face identification tasks. The goal is to generalize them to other tasks such as the verification of unseen identities in the training set. The DeepID features are taken from the last hidden layer neuron activations of deep CNN and learned as classifiers to recognize about 10,000 face identities in the training set. Specifically, each CNN is configured to take a face patch as input and extracts local low-level features in the bottom layers, thus it keeps reducing the neuron numbers along the feature extraction hierarchy while gradually more global and high-level features are formed in the top layers.

However, it is impractical to collect a huge number of labeled images and train a large network on the best feature representation for every task, and then generalize the model for other different tasks. This challenge has motivated Azizpour et al. [230] to assess the relation between choosing and learning CNNs representations and the performance on a diverse set of visual recognition tasks. Particularly, they have investigated how altering the parameters in the CNN network architecture impacts the representation ability to specialize and generalize, and the effect of fine-tuning a generic network towards a particular task. Based on extensive experimental results and using many standard image datasets, they indicated that increasing the specialization will increase the performance on the target task, but this may affect the ability to generalize it to other tasks.

Another successful advancement using deep neural networks has been recently proposed and announced as a promising solution for automatic image annotation. Karpathy and Fei-Fei [231] have introduced a model that generates free-form natural language descriptions of image regions. The proposed model leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between text and visual data. It is based on a novel combination of the CNN over image regions, bidirectional recurrent neural networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Finally, the constructed model describes the RNN architecture that uses the inferred alignments to learn generating novel descriptions of image regions. The overall procedure of their novel method for image annotation using deep learning is demonstrated in Fig. 6.

After learning the model on image patterns and descriptions, it has been applied on unseen images in different large datasets. The results showed a high capability in identifying image objects and actions (i.e. events) with a relatively high accuracy, and sometimes roughly near the human perception capabilities. Fig. 7 shows a sample set of automatically annotated images. This proposed deep architecture has provided robust and qualitative results that may make annotating and searching for billions of images better and easier than previous manual or semi-automatic annotation methods which often poorly describe the image semantics. Table 5 summarizes the learning methods discussed in this section.

2.5. Relevance feedback

The relevance feedback (RF) process has been adopted in many CBIR systems and noticeably improves the search accuracy and user satisfaction. Generally, users submit a query and seeking for relevant images with high satisfaction, but they have a difficulty in formulating what they actually need by the textual forms even with the existence of user-friendly interfaces. Therefore, RF algorithms provide an interaction between the CBIR system and users to indicate whether the returned results are relevant or not.

Basically, the RF procedure includes these main steps: (1) the user initiates a sample image query, (2) the system processes the query and returns an initial set of images as a result, (3) the user

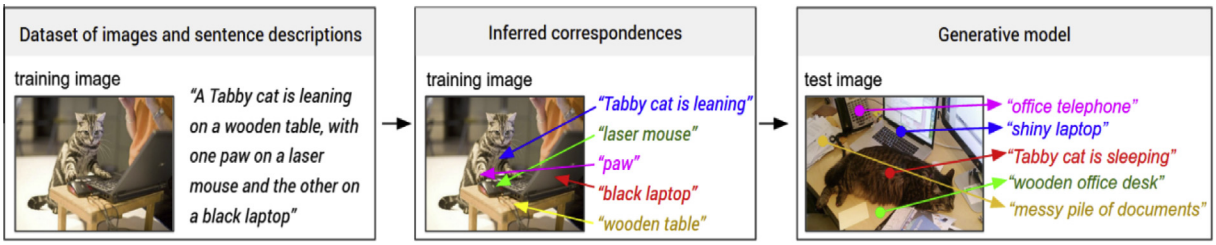

Fig. 6. Overview of the proposed model in [231].

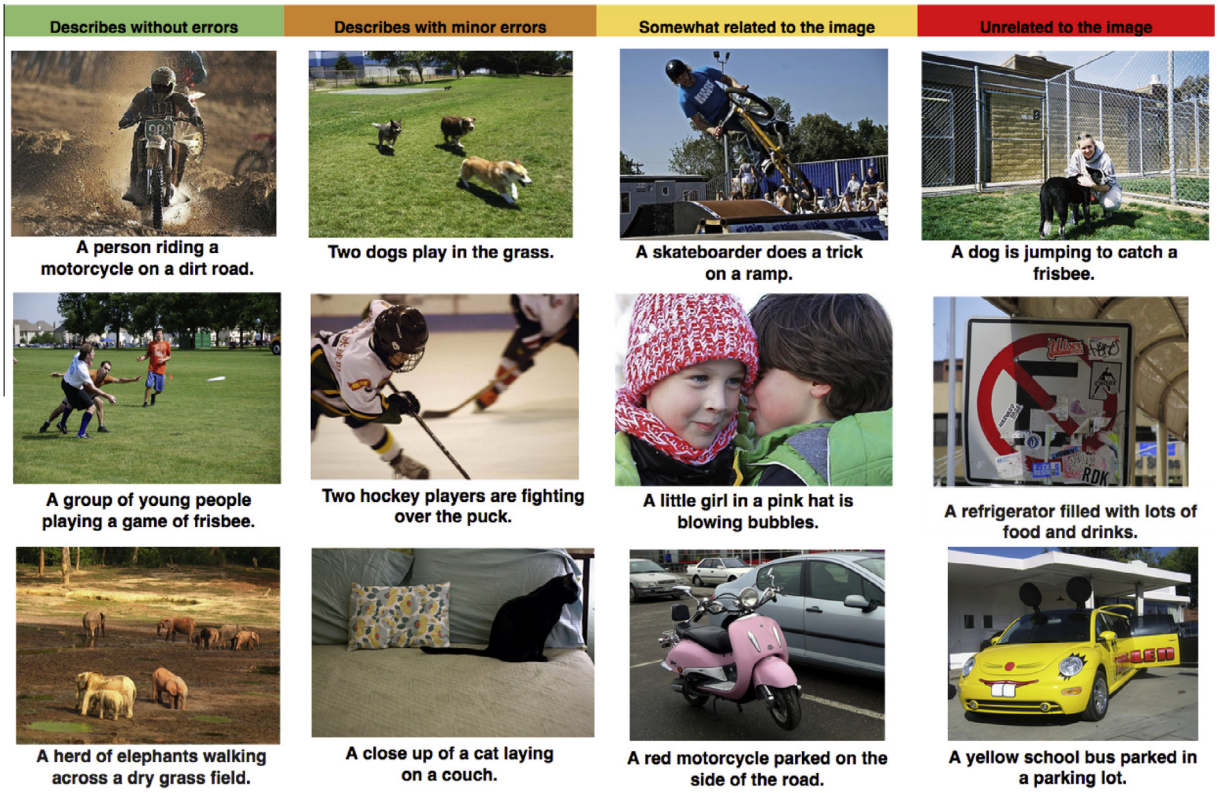

Fig. 7. A selection of evaluation results that grouped by human rating [232].

Table 5
The main characteristics of learning methods.

Methods	Main attributes	Limitations	Examples	Semantic contributions
Unsupervised Learning (Clustering)	– No prior knowledge about image semantics and categories – Incremental and iterative – Offers hard/soft and hierarchical/flat clustering (application-dependant)	– Trade-off between the number of clusters and the clustering performance – Scalability and overfitting	– *k*-means – GMM – Fuzzy – Constraint-based – Concept-based	Learns the underlying semantic properties and similarities of data input and generalizes the process to deal with new unseen images
Supervised Learning (Classification)	– Based on predefined image classes – Offers generative and discriminative classification	– The classification accuracy is directly proportional with the number of training data points	– SVM – Bayesian – Ensemble – ANNs – Regression	The pre-defined semantic class hierarchy reflects in the semantics of the human's subject, so it is flexible and intuitive
Deep Learning	– Automatically learns features and complex tasks from data at multiple levels of abstracts – Generative learning model – (Un)Supervised pre-training which uses labeled/unlabeled training data efficiently	– High computation cost – Complex architectures and many parameters need to be estimated and controlled – Theoretically, unclear network structures that are seen as a black box that need to be investigated practically	– CNNs – Deep Neural Networks – Deep Belief Networks – Boltzmann Machines	– Models high-level semantics in data by utilizing multiple feature representations, similarity measures, and non-linear transformations – Provide an efficient and automatic image annotation (for objects and actions)

indicates some of images as relevant/irrelevant (i.e. labeled as positive/negative samples), (4) the system updates/reweights its representations and metrics based on the user feedback, (5) the system returns a revised set of images that are more relevant to the user query, and finally (6) the system might perform these steps in many iterations until the user gets satisfied. However, most of users prefer to get desired images with a minimal intervention rather than performing many RF iterations. In addition, the query image submitted does not exactly reflect the intended semantics and it is difficult to formulate them in a textual/visual query. As a result, the retrieval system might ignore some relevant images while updating the retrieved images according to the user feedback. Therefore, the research efforts have proposed a variety of RF algorithms in order to solve these problems and improve the CBIR accuracy with minimum performance degradation.

2.5.1. RF categories

It is important to learn the user feedback to effectively enable an automatic refinement on image results with a minimal user intervention. The RF approaches can be broadly categorized into two groups [18]: (1) short-term RF learning, which considers only the current feedback session and ignores previous data from other users and (2) long-term RF learning, which records and consider feedback knowledge from different users based on their feedback. Since the short-term RF approaches just rely on the current feedback in labeling the image samples as positive/negative, they might result in a considerable loss of relevant data. Here, we present some important RF techniques that have been recently proposed in the CBIR domain.

2.5.1.1. SVM-based RF. It is one of the most widely used approaches in the field of CBIR. Despite its success, it considers the positive and negative feedbacks equally which is not robust base since each sample has distinct properties. In addition, most of SVM-based RF approaches do not consider unlabeled samples which may yield a weak classifier during the refinement process. Recently, Bian and Tao [233] have proposed a new related SVM-based RF learning, referred as the biased discriminative Euclidean embedding (BDEE), which parameterizes the image samples in the original high-dimensional space to discover the intrinsic coordinate of low-level visual features, i.e. the manifold structure. This manifold regularization-based structure is merged with BDEE to form a semi-supervised BDEE which takes into account the unlabeled samples. The key advantage of the BDEE is the ability to model both the intraclass geometry and the interclass discrimination which avoids the under-sampled problem. Zhang et al. [234] have also proposed another SVM-based RF scheme, namely the biased maximum margin analysis (BMMA) and semi-supervised BMMA (SemiBMMA), to utilize the information of unlabeled samples and integrate the distinct properties of feedbacks. The BMMA distinguishes the positive feedbacks from the negative ones based on a local analysis, whereas the SemiBMMA can effectively integrate the information of unlabeled samples by employing the Laplacian regularizer in the BMMA.

2.5.1.2. Biased discriminant analysis (BDA)-based RF [235]. It is a promising approach where both positive and negative feedbacks are not deemed equivalent. In the BDA model, all positive samples are considered alike and required to stay away from the negative feedbacks. However, users usually give a small number of feedback samples. Moreover, the performance of the BDA-based RF for CBIR applications suffers from (1) the singular problem of the positive intraclass scatter and (2) the Gaussian distribution assumption for positive samples. To avoid these intrinsic problems, Zhang et al. [236] have recently proposed another approach called the generalized BDA (GBDA) for various CBIR tasks. The GBDA first avoids the singular problem by adopting the differential scatter

discriminant criterion (DSDC), and it redesigns the between-class scatter by the nearest neighbor approach to handle the Gaussian distribution assumption. The GBDA integrates the locality preserving principle to alleviate the overfitting problem; therefore, a locally consistent and smooth transform can also be learned. Based on extensive experiments, the authors have shown that the GBDA can substantially outperform the original BDA and other related SVM-based RF algorithms.

2.5.1.3. Case-based reasoning RF. A frequent adding of images to the CBIR system may result in degrading the performance of long-term learning RF methods. Additionally, the recorded feedbacks in such methods are often sparse. To treat these problems, Rashedi et al. [237] have proposed a long term learning method by adopting the case-based reasoning (CBR), which is called the case-based long term learning (CB-LTL). The CBR technique uses a specific knowledge of previously experienced problems (i.e. cases) to solve a new problem by finding the similar past case and reusing it in the new problem situation. The CB-LTL method has two stages: learning and reasoning. In the learning stage, the extracted information from retrieval sessions is saved as cases. In the reasoning stage, the information about cases is utilized to improve the results of retrieval sessions. The main components of the CB-LTL method are: the 'key of query' which represents the user desire, the 'trigger function' which is used to find a similar case with a query, and the 'semantic frame' which is a structure for saving cases. In the proposed method, the cases are recorded in the case knowledge base using both low-level and high-level features. The information of the relevance feedback and short-term learning are employed as high-level features.

2.5.1.4. Genetic-based RF. To consider the user preferences and subjectivity during the retrieval process, Lai and Chen [238] have proposed a user-oriented mechanism based on the interactive genetic algorithm (IGA) in the CBIR domain. The color (i.e. mean values, standard deviation, and image bitmaps) and texture (i.e. entropy based on the COM and the EH) features are used as image features. To reduce the gap between the returned results and user expectation, the IGA is employed to help users in identifying the images that are most relevant to their needs. Another RF approach based on the genetic programming has been proposed by Ferreira et al. [239], which considers both relevant and irrelevant images that indicated by the user. The proposed method has employed color, texture, and shape as image descriptors to represent the content of database images, and showed its efficiency compared with other RF methods.

2.5.1.5. Others. Bulo et al. [240] have developed a novel approach based on the random walker algorithm introduced in the domain of interactive image segmentation. The key idea is that the relevant and non-relevant images labeled by the user at every feedback round are treated as seed nodes for the random walker problem. The ranking score is computed for each unlabeled image as the probability that a random walker starting from that image will reach a relevant seed before encountering a non-relevant one. Most of existing RF-based CBIR methods usually produce the refined search results by a number of iterative feedbacks, but this is impractical and inefficient in the real-world applications which process large-scale image datasets. Su et al. [241] have proposed a novel breadth-first search (BFS)-based KNN method, the navigation-pattern-based relevance feedback (NPRF). It achieves a high efficiency and effectiveness for CBIR applied on large-scale image datasets. In terms of efficiency, the iterations of feedback are reduced fundamentally using the navigation patterns discovered from the user query log. The NPRF-Search uses the navigation patterns obtained and three kinds of query refinement: query reweighting (QR), query point movement (QPM), and query expansion (QEX), to effectively converge the search space toward the user intention. Based on the NPRF, a high quality of image retrieval can be achieved in a small number of feedbacks. As important complementary methods to RF methods, visual analytics and interactive information retrieval are discussed in the following section.

2.5.2. Visual analytics and visualization

The current CBIR systems present the retrieved images as a ranked list which does not provide an explicit description about the intrinsic properties of retrieval, similarity, ranking, and the relationship between features represented in the data space and images ranked in the result set. Consequently, this may yield some relevant images ranked at low level in the result set which limits the user ability to explicitly identify these relevant images during the refinement process. This problem is very close to the notion of the information overload problem of losing data which may be (1) irrelevant to the current task, or (2) processed and presented in an inappropriate way [242]. Visual analytics (VA) is an emerging approach which is employed to deal with these problems and improve the CBIR accuracy, user satisfaction, and performance. Generally, the VA is a multidisciplinary approach which is beneficial in several focus areas such as analytical reasoning, visual representation and interaction, data representation and transformation, and supports the production, presentation, and dissemination of analytical results [243].

VA methods play an integral role in bridging the semantic gap and assisting users to make their queries and selections clear and adaptable to changes of their intentions. Hiroike et al. [244] have developed a system that presents the results of a CBIR system as a dynamical scatter diagram which includes image thumbnails. The developed user interface provides different similarity-based transformations from a high-dimensional feature space to a 3D space that give different coordinate systems in the visualization space. Rodrigues et al. [245] have combined CBIR and VA methodologies to improve the possibility of understanding the metric space that supports the similarity queries which allows fine tuning and advanced use of distance functions, feature extraction, and metric data structures. Additionally, the system offers several functionalities such as browsing a dataset in a tabular format, data visualization, and saving similarity queries into visualization workspaces. Recently, Kumar et al. [246] have developed a tool, namely the visual analytics for medical image retrieval (VAMIR), which enables a guided visual exploration of large features in the search space. The VAMIR utilizes the query image as a point-of-reference for dynamic querying, automatic feature selection, and dynamic modification on the visualizations. This tool interactively adjusts search parameters through an analytical approach, and hence it is useful for scientific and education applications of medical CBIR systems. The VA is one of the visualization aspects, but the visualization capability has wider directions in the CBIR domain. Image interpretations by both human and machine are very related to the human–machine interaction mechanisms conducted during the retrieval session. CBIR systems' GUIs should have the ability to visualize the relationship and similarities between the user query and returned images as well as between database images themselves. Tory and Moller [247] have discussed how human factors significantly contribute to the visualization process and its influences on the design and evaluation of perception-based visualization tools. The contributions of human factors include, but not limited to, utilizing theories of perception, designing a system fits a particular task or human capability, and user involvement in design and evaluation. Data visualization contributes in bridging several CBIR gaps related to the semantic ones such as content, feature, performance, and usability gaps [248]. The content gap characterizes

the human understanding of images, including the general context where the system may be used. The feature gap describes the granularity of image structure and visual details which recognized and processed by the system in addition to the automation of feature extraction. The performance gap characterizes system implementation, integration with other systems, search optimization, and quantitative evaluation. Finally, the usability gap which is very close to the user cognition in terms of user ability in using textual/visual queries, understanding the similarities and measures between images, providing easy and understandable relevance feedback, refining the query results, and learning user preferences.

Wilson [249] has provided a comprehensive framework about the elements and factors that make up different search user interface (SUI) designs. The SUI includes a multidisciplinary vision goes beyond simply submitting a query and displaying results. It broadly considers browsing interfaces, sense-making problems, and scenarios of learning and decision making. They state that the SUI elements could be broadly grouped into four categories: (1) input features that easily allow users to express what they are looking for; (2) control features that help users to modify, refine, limit, or expand their input; (3) information features that provide results; and (4) personalisable features that relate to users themselves and their previous interactions.

However, a powerful retrieval algorithm can be rendered useless or unused by a frustrating SUI and poorly managed data. As a result, the retrieval community realizes that more specific queries lead to more relevant results. The interactive information retrieval (IIR) deals with these aspects and others; including refining and improving a query. The IIR provides an evaluation of the relationship between human–system and human–information interactions [250]. Ingwersen and Järvelin [251] have integrated both information seeking (IS) and information retrieval (IR) into one notion called the IS&R which defines a broader cognitive viewpoint of humans involvement and interaction in the retrieval relevance evaluation by means of information sources and IR systems. Recently, Kumar et al. [252] have designed a CBIR user interface to assist the user interpretation of retrieved 3D and multi-D medical images. It provides multiple views of volumetric and multi-modality images, abstractions summarize complex data, tools for result refinements, and visualization for the similarities between images. Table 6 summarizes the RF and VA methods and their characteristics.

Table 6
The main characteristics of relevance feedback methods.

Methods	Main attributes	Limitations	Examples/ Variants	Semantic contributions
SVM-based	– Model the intraclass geometry and interclass discrimination – The most ambiguous images are filtered or labeled by the user	– Most of methods consider positive and negative feedbacks equally, and do not consider unlabeled samples – Low accuracy due to the small sample size	– BDEE [233] – BMMA [234]	An active learning which separates the relevant from the non-relevant images in a high dimensional feature space
BDA-based	– Positive and negative feedbacks are not deemed equivalent – A small number of feedback samples provided by users	The positive intraclass scatter and the Gaussian distribution assumption for positive samples	– BDA [235] – GBDA [236]	The original input space is nonlinearly mapped to an arbitrarily high dimensional feature space
Case-based	Based on a specific knowledge of previously experienced problems (cases) to solve a new problem	The accuracy depends on the amount and the similarity of the saved courses	CB-LTL [237]	– 'Key of query' represents the user desire – 'Trigger function' finds a similar case with a query – 'Semantic frame' is a structure for saving cases
Genetic-based	– User-oriented – Considers both relevant/irrelevant samples	– Complex similarity functions defined for feature vectors – Computationally expensive	IGA-RF [238] GP-RF [239]	Provide more non-linear combination among the similarity values to express the user needs
Navigation-Pattern	– Need few iterations due to the navigation patterns discovered from user query log	– Scalability	NPRF [241]	Multiple and hierarchical query refinement to converge the search space toward the user's intention
Visual Analytics	– Interactive user interface – Dynamic querying and ranking – Provide more descriptions about the retrieval process	– Need powerful equipment – Not all users prefer to iteratively interact through GUI visualizations	VAMIR [243] Scatter Diagram [245] Visualization fusion [246]	Enable users to adjust the retrieval parameters based on their semantics through an interactive GUI

2.6. Distance measures

In general, the structure of feature vectors selected determines the type of distance measure that will be used to compare their similarity. The applied distance measure mathematically indicates the similarity between the query and each image/region/object in the database. To achieve more accurate retrieval and better performance, the CBIR system should employ an effective similarity matching measure which accurately characterizes and quantifies the perceptual similarities. Despite the success of utilizing the common distance measures in the literature, finding an adequate and robust distance measure is still one of the challenging issues in the field of CBIR. In the following sections, we discuss the common distance measures, graph-based similarity measures, and distance metric learning.

2.6.1. Types of distance measures

The most common distance measures used for similarity matching in the CBIR domain are:

2.6.1.1. Minkowski distance. It is considered as the most widely used metric for measuring the similarity in CBIR systems. Given

two images X and Y that represented in the data space by two n-dimensional vectors $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, respectively. The Minkowski distance between X and Y, $d(X,Y)$, is defined as:

$$d(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{1/r}, \quad (4)$$

where r is the norm factor for Minkowski distance, and $r \geqslant 1$. When $r = 1$, $r = 2$, and $r = \infty$, it becomes the well-known Manhattan (i.e. L_1), Euclidean (i.e. L_2), and Chebyshev (i.e. L_∞) distances, respectively.

2.6.1.2. Mahalanobis distance. Given point A and distribution B, the Mahalanobis measures the distance between A and B by computing how many standard deviations away A is from the mean of B. Let the covariance matrix M, and two images X and Y that represented in the data space by two n-dimensional vectors $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, respectively. The Mahalanobis distance between X and Y is defined as:

$$d(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r S^{-1}\right)^{1/r}, \quad (5)$$

if $r = 2$ and the covariance matrix S is identity matrix then it is equivalent to the Euclidean distance, but if S is diagonal matrix then it is equivalent to the normalized Euclidean distance.

2.6.1.3. Cosine distance. Given two images X and Y that represented in the data space by two n-dimensional vectors, the distance is given by the angle between vectors using the dot product and magnitude as:

$$d(X,Y) = 1 - \cos\theta = 1 - \frac{X \cdot Y}{\|X\| \cdot \|Y\|}. \quad (6)$$

2.6.1.4. Hamming distance. Given a finite data space F with n elements, the Hamming distance $d(x,y)$ between two vectors $x, y \in F^{(n)}$ is the number of coefficients in which they differ, or can be interpreted as the minimal number of edges in a path connecting two vertices of n-dimensional space. In the CBIR system, the hamming distance used to compute the dissimilarity between the feature vectors that represent database images and query image. The fuzzy Hamming distance (D) is an extension of Hamming distance for vectors with real values.

Given the real values x and y, the difference degree between x and y, modulated by $\alpha > 0$, denoted by $d_\alpha(x,y)$ is defined as [253]:

$$d_\alpha(x,y) = 1 - e^{-\alpha(x-y)^2}, \quad (7)$$

and the parameter $\alpha \geqslant 0$ modulates the difference degree in α since that for the same value of $|x-y|$ different values of α will result in different values of $d_\alpha(x,y)$. The membership function d_α defined in (6) has the following properties:

1. $0 \leqslant d_\alpha(x,y) < 1$ with equality if $x = y$;
2. $d_\alpha(x,y) = d_\alpha(y,x)$;
3. for $x = a \pm c$, $d_\alpha(x,a) = e^{-c2}$; and
4. $d_\alpha(x,y) = d_\alpha(0, |x-y|)$.

Using the notion of the difference degree defined above, the difference fuzzy set $D(x,y)$ for two vectors x and y is defined as follows [253]: let x and y be two n dimensional real vectors and x_i, y_i denote their corresponding ith component. The difference degree between x and y along the component i, modulated by the parameter α is $d_\alpha(x_i, y_i)$. The difference fuzzy set corresponding to $d_\alpha(x_i, y_i)$ is $D_\alpha(x,y)$ with membership $\mu D_\alpha(x,y)$ function:$\{1, \ldots, n\} \to [0\ 1]$ is *given by*:

$$\mu D_\alpha(x,y)(i) = d_\alpha(x_i, y_i), \quad (8)$$

which is the degree to which vectors x and y are different along their ith component.

2.6.1.5. Earth Mover's distance. The EMD [254] is based on the transportation problem from linear optimization which targets the minimal cost that can be paid to transform one distribution into the other. For image retrieval, this idea is combined with a representation scheme of distributions which is based on vector quantization for measuring perceptual similarity. This can be formalized in a linear programming problem as follows: Let $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$ is the first signature with m clusters, where p_i is the cluster representative and w_{pi} is the cluster weight; and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$ is the second signature with n clusters; and $D = [d_{ij}]$ is the matrix of ground distance where d_{ij} is the ground distance between clusters p_i and q_j . To compute a flow $F = [f_{ij}]$, where f_{ij} is the flow between p_i and q_j, that minimizes the overall cost:

$$\mathrm{WORK}(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}, \quad (9)$$

this is subject to the following constraints:

(1) $f_{ij} \geqslant 0$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$.
(2) $\sum_{i=1}^{n} f_{ij} \leqslant w_{p_i}$, $1 \leqslant i \leqslant m$.
(3) $\sum_{j=1}^{m} f_{ij} \leqslant w_{qj}$, $1 \leqslant j \leqslant n$.
(4) $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{qj}\right)$.

Constraint (1) allows moving supplies in one way from P to Q; Constraint (2) limits the amount of supplies that can be sent to their weights by the clusters in P; Constraint (3) limits the clusters in Q to receive no more supplies than their weights; and constraint (4) forces to move the maximum possible amount of supplies that called the total flow. Once the optimal flow F is found and the transportation problem is solved, the EMD is defined as the resulting work normalized by the total flow as follows [254]:

$$\mathrm{EMD}(P,Q) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \Big/ \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}. \quad (10)$$

The EMD is more robust than histogram-based matching techniques and has many advantages over other definitions of distribution distances. First, the EMD applies to signatures which subsume certain histograms. This holds the advantages of signatures compactness and flexibility as well as the benefit of handling these variable-size structures by a distance measure. Second, the cost of moving earth properly reflects the notion of nearness without the existence of quantization problems of most other measures. Third, the EMD offers a partial matching which is important, for instance, to deal with clutters and occlusions in image retrieval applications. Fourth, if the ground distance is a metric and with equal total weights of two signatures, the EMD allows endowing image spaces with a metric structure [254].

2.6.1.6. Kullback–Leibler and Jeffrey divergence distance. Based on the information theory, the K–L divergence measures how inefficient on average it would be to code one histogram using the other one as code-book. Given two histograms $H = \{h_i\}$ and $K = \{k_i\}$, where h_i and k_i are the histogram bins, the Kullback–Leibler (K–L) divergence is defined as follows [255]:

$$d_{KL}(H,K) = \sum_{i=1} h_i \log(h_{i/k_i}). \quad (11)$$

However, the K–L divergence is sensitive to histogram binning and non-symmetric. The empirically derived Jeffrey divergence is a modification of K–L divergence that is numerically symmetric, stable, and robust to the noise and the size of histogram bins. This distance measures how unlikely it is that one distribution was drawn from the population represented by another one and defined as [255]:

$$d_J(H,K) = \sum_i (h_i \log(h_i/m_i) + k_i \log(k_i/m_i)), \quad (12)$$

where $m_i = (h_i + k_i)/2$, and for χ^2 Statistics:

$$d_\chi^2(H,K) = \sum_i (h_i - m_i)^2/m_i. \quad (13)$$

Table 7 summarizes the types of distance measures and lists the main characteristics of each type.

2.6.2. Graph-based similarity measures

With the rabidly increasing amounts of graph-based representations and features in the domain of image retrieval, providing a robust graph-based similarity measures is an important research problem. Given a graph database with n graphs, i.e. $G = \{g_1, g_2, g_3, \ldots, g_n\}$, and a query graph q, the similarity graph search can be defined as: find all graphs g_i in G such that g_i are similar to q within a predefined threshold based on certain similarity measures. Many graph-based similarity measures have been proposed, including the distance metrics based on maximal common subgraphs [256], maximal/minimal common subgraph [257,258], and graph edit distance (GED) [259]. The most widely applied method for the graph similarity measurement is the GED. Basically, the GED computes the cost of elementary operations; i.e. node substitution, node insertion/deletion, edge insertion/deletion. The edit distance between two graphs is the minimal cost taken over all of these operations. The GED can be applied to any type of graphs and robust in the presence of noise and errors in the database. Since the GED is NP-hard and has an exponential computational complexity, Zeng et al. [260] approximate the edit distance by the notion of star representation for full and subgraphs structures. It obtains the lower and upper bounds of the GED between two graphs in polynomial time.

2.6.3. Distance metric learning

The distance metric learning is a very important factor in determining the quality of a CBIR system in terms of accuracy and performance. The aim is to learn a distance metric for the input data space from a given collection of pair of similar/dissimilar points that preserves the distance relation among training data. Therefore, many studies have been proposed and proved that the learned metric can significantly improve the performance in classification, clustering, and retrieval tasks [261]. Distance metric learning can be roughly divided into three categories according to the learning algorithm [261,262]: (1) unsupervised learning methods such as in [191,192], which do not rely on the labeled information to learn the distance metric and usually used for unsupervised learning tasks such as clustering and dimension reduction; (2) supervised learning methods such as in [263,264], which learns the distance metric from the training data with class labels and keeps similar data points in the same class; and (3) semi-supervised learning methods such as in [265], which use both labeled and unlabeled data which is beneficial when only limited labeled data available.

Zhang et al. [266] have proposed a new metric learning under the transfer learning setting where some source and target tasks

Table 7
The most commonly used distance measures in CBIR applications.

Measures	Main attributes	Limitations	Equation	Usage/Domains
Manhattan-L_1	Less affected by outliers and therefore noise in high dimensional data	Yields many false negatives because of ignoring the neighboring bins, and gives near and far distant components the same weighting	Eq. (4) ($r = 1$)	– Computes the dissimilarity between color images – e.g. fuzzy clustering
Eucledian-L_2	Allows normalized and weighted features	– Sensitive to the sample topology – Does not compensate for correlated variables	Eq. (4) ($r = 2$)	The most commonly used method, e.g. k-means clustering
Chebyshev-L_∞	– Maximum value distance – Induced by the supremum norm/uniform norms	Does not consider the similarity between different but related histogram bins	Eq. (4) ($r = 3$)	Computes absolute differences between coordinates of a pair of objects, e.g. fuzzy c-means clustering
Mahalanobis	– Quadratic metric – Incorporates both variances and covariances	Computation cost grows quadratically with the number of features	Eq. (5)	Improves classification by exploiting the data structure in the space
Cosine	Efficient to evaluate as only the non-zero dimensions considered	Not invariant to shifts in input	Eq. (6)	Efficient for sparse vectors
Hamming	Efficient in preserving the similarity structure of data	Counts only exact matches	Eqs. (7) and (8)	– Identifies the nearest neighbor relationships – e.g. Image compression, and vector quantization
EMD	– Signature-based metric – The ability to cluster pixels in the feature space – Allow partial matching	Not suitable for global histograms (few bins invalidate the ground distances, while many bins degrades the speed)	Eq. (10)	– Useful metric between signatures in different spaces – Robust against clutters and occlusions – Efficient for clustering
K-L divergence	– Asymmetric – Non-negative	– Sensitive to histogram binning	Eq. (11)	Computes dissimilarity between distributions, e.g. texture-based classification
J divergence	– Symmetric – Stable and robust to noise and histogram binning	– Difficult multivariate estimation for limited samples	Eq. (12)	Computes the dissimilarity between distributions, e.g. texture-based classification

are available, which referred as the transfer metric learning (TML). Unlike the conventional distance learning methods, the TML is based on the convex formulation for multitask metric learning. It models the task relationships in the form of task covariance matrix which can model positive, negative and zero task correlations. Additionally, the TML learns the metric and task covariances between the source and target tasks under a unified convex formulation. However, there are a large number of classes in many image classification problems; so learning a global metric to improve the image-to-class (I2C) distance may not be sufficient. Accordingly, Wang et al. [262] have proposed a method for distance metric learning to improve the performance of I2C distance by learning different Mahalanobis matrices for every class. This metric can preserve the discriminative information for different classes during the learning procedure, thus provides better performance than learning only a global metric. Consequently, the per-class metric learning developed corrects the class imbalance problems where there are more training images in some classes than others.

3. Image datasets

A wide range of image benchmarking datasets have been used in the CBIR domain that range from a small set of crawled images by human to a large number of images. However, the selection of image dataset is crucial and mainly depends on the application, adopted algorithm, and problems addressed. For instance, CBIR systems developed for object detection test and validate the retrieval accuracy and performance using a dataset of object images (i.e. the images contain shaped objects such as human, animals, and natural items). The majority of image datasets usually contains various categories and each category (i.e. image class) has similar semantics as shown in Fig. 8. For example, Wang's image dataset consists of 1000 colorful images with 10 different categories and each category contains 100 images. This semantic categorization of the dataset is determined by the authors, which reflects the human perception of image similarity.

Often, image datasets are manually grouped and annotated into semantic classes so some of limitations should be considered before choosing the image dataset. Firstly, some images are challenging and can belong to many classes due to the substantial variations in scale and viewpoint. For instance, monuments image in Fig. 8 could be assigned to three different classes: monuments, buildings, and city. Secondly, some images belong to different classes, even they have similar semantic contents (e.g. images belong to the beach and mountains classes may have very similar contents such as sky, stones, and water). Finally, datasets with image labels (or tags) are not accurately annotated to reflect high-level concepts which impact the performance and accuracy of the retrieval process. However, many CBIR systems developed for different problem domains, thus using small image datasets might be insufficient to evaluate the performance of large-scale image retrieval systems. Additionally, learning-based CBIR systems require a sufficient amount of images to effectively train and test image concepts during the classification and retrieval processes. Table 8 lists the common image databases used in CBIR systems, including roughly the number of images and classes, attributes, and some special versions.

Corel dataset is one of the most used image databases, which contains a large amount of images with various semantic contents and groups [267]. ImageNet [268] is another image database that widely used in the CBIR domain. It is organized according to the WordNet hierarchy (currently only the nouns) where each node of the hierarchy is depicted by hundreds or thousands of images with an average of over five hundred images per node. Other image datasets are more suitable for certain applications and tasks such as Oxford and Paris (part of Flicker dataset) that are suitable for systems deal with city and landmarks semantics.

Additionally, several benchmarking and evaluation competitions have been developed in a wide range of disciplines. ImageCLEF [281] provides support for the evaluation in the field of visual information and analysis, indexing, classification, and retrieval. It has been launched in 2003 with open participation from academic and industrial research groups worldwide from various global communities such as the information retrieval, computer vision, pattern recognition, and medical informatics. Over the last years, ImageCLEF tasks introduced became more challenging due to the emerging trends and the growing size of image collections. Datasets and tools provided by ImageCLEF are all freely available, and the results of its annual competition are published.

Pascal-VOC [280] is a visual classes' recognition challenge which provides publicly available datasets of images together with the ground truth annotation and evaluation software. It has been

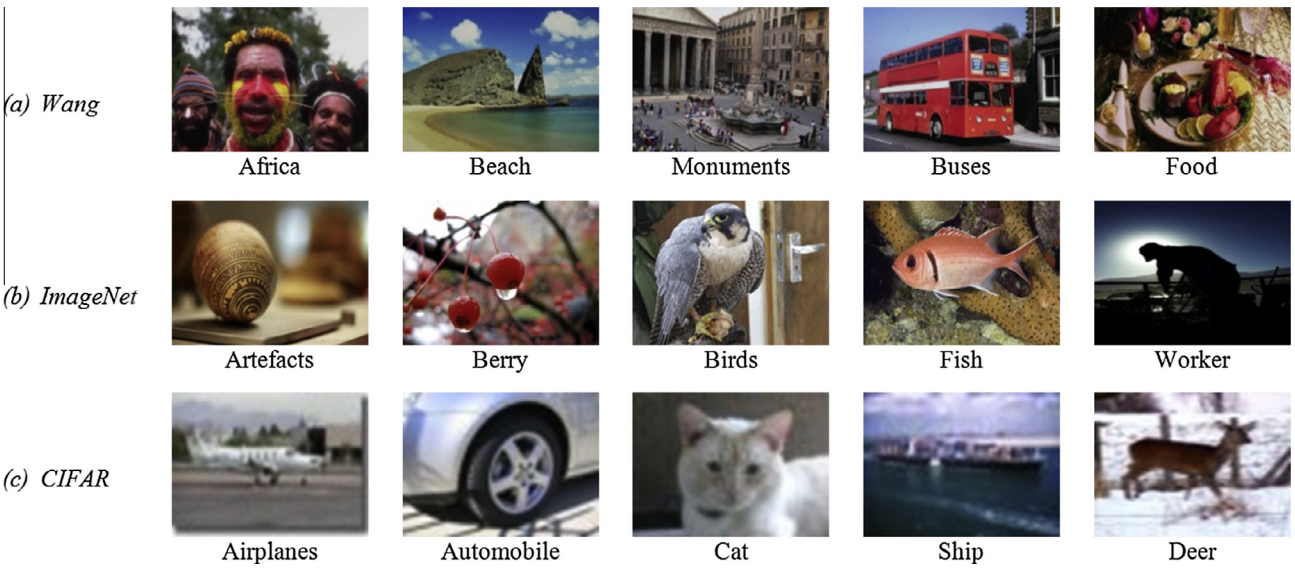

Fig. 8. Samples of classes from Wang, ImageNet, and CIFAR image datasets.

Table 8
Some commonly used image datasets in the CBIR domain.

Datasets	# Images[a]	Some versions	# Classes[a]	Annotated/Labeled
ImageNet [268]	>14.2 M	ILSVRC	>21.8 K tags	✓
CIFAR [269]	60 K	CIFAR-10	10	✓
		CIFAR-100	100	
CALTECH [270,271]	>9 K	Caltech-101	102	✓
	>30.6 K	Caltech-256	257	
FLICKR [272]	>7 B	YFCC (100 M)	>100 concepts	✓
		MIR-Flickr (1 M)		
		Oxford (>45 K)		
WANG [273]	1 K	SIMPLIcity	10	
	10 K	WBIIS		
LabelMe [274]	>30 K	N/A	183	✓
TinyImage [275]	>79 M	N/A	>75 K tags	✓
SUN [276]	>131 K	Scene397	908 Scenes > 4.4 K Objects	✓
		SUN2012		
NORB [277]	>29 K	N/A	6	✓
NUS-WIDE [278]	>269.6 K	NUS-WIDE-LITE	81	✓
		NUS-WIDE-OBJECT		
		NUS-WIDE-SCENE		
SUN-Attribute [279]	>14 K	N/A	>700 tags	✓
PASCAL- VOC2007 [280]	>9.9	VOC2005 – VOC2012	20	✓

[a] K (thousands), M (millions), B (billions), N/A (Not-Applicable).

launched in 2005 and divided into five challenges: classification, detection, segmentation, action classification, and person layout. The competition has certainly contributed in the computer vision community, especially in the performance evaluation of different algorithms and training/testing datasets. ILSVRC [282] is another annual competition that evaluates algorithms for object detection and image classification especially in large-scale image retrieval. It has been launched in 2010 and allows researchers to compare the progress in detection across a wider variety of objects, and measures the computer vision progress for large-scale image indexing, retrieval, and annotation. Like other competitions, the ILSVRC provides datasets, development kits, and statistics to the public.

4. Performance evaluation

The performance evaluation of CBIR systems is usually handled by predefined system criteria rather than relying on the user intervention which is inaccurate, human-subjective, and time consuming. However, there is no single standard criterion to evaluate the accuracy and performance of CBIR systems. Instead, there are set of common and trusted measures that have been used in the literature. Choosing a suitable measure generally depends on the CBIR method, problem domain, and the algorithm itself. The following are the most commonly used evaluation metrics:

4.1. Precision and recall

Precision (P) and recall (R) are usually used for performance evaluation in CBIR systems. Precision (P_k) is the ratio of the number of relevant images (N_R) within the first k results to the number of total retrieved images (N), and defined as:

$$P_k = \frac{N_R}{N} = \frac{(relevant\ images \cap retrieved\ images)}{retrieved\ images} \tag{14}$$

Recall (R_k) is the ratio of the number of relevant images within the first k results to the number of total relevant images (N_{RV}), and defined as:

$$R_k = \frac{N_R}{N_{RV}} = \frac{(relevant \cap retrieved)}{relevant\ images}. \tag{15}$$

4.2. F-measure

The use of precision or recall alone is generally considered insufficient for accuracy evaluation thus combining them into one measure is possible. The F-measure (or $F1$-score) is the harmonic mean of precision and recall, and defined as:

$$F_1 = 2 \cdot \frac{P_k \cdot R_k}{P_k + R_k},\ \text{and the generalized form for positive } \beta \text{ is :}$$

$$F_\beta = (1+\beta^2) \cdot \frac{P_k \cdot R_k}{(\beta^2 \cdot P_k) + R_k}.$$

4.3. Average-precision

It is another global estimation measure for the retrieval performance using a single value known as the average precision (AP). For a single query q, the AP is the mean over the precision values at each relevant image:

$$AP_q = \frac{1}{N_{RV}} \sum_{k=1}^{N_{VR}} P_k(R_k). \tag{16}$$

4.4. Mean average precision:

Given a set of queries N_Q, the mean average precision (MAP) is the mean of the average precision scores over all queries, and defined as:

$$\mathrm{MAP}_Q = \frac{1}{N_Q} \sum_{q=1}^{N_Q} AP_q. \tag{17}$$

4.5. Precision–Recall curve

In rank-based retrieval, appropriate sets of retrieved images are given by the top k retrieved images. For each set, precision and recall values can be demonstrated by a graphical curve. Precision–recall curves have a distinctive sawtooth shape, i.e. if the $(k+1)$th image retrieved is relevant, then both the precision and recall increase and the curve jags up to the right, otherwise the recall is the same as for the top k images but the precision is

dropped. It is often useful to remove the sawtooth shape of curves and the standard way to do this is by the interpolated precision-recall, i.e. the interpolated precision at a certain recall level r is defined as the highest precision found for any recall level $r' \geqslant r$:

$$P_{\text{interp}}(r) = \max_{r' \geqslant r} p(r'). \tag{18}$$

The advantage is that users would be prepared to look at few more images if this would increase the percentage of the viewed set that was relevant, i.e. if the precision of the larger set is higher. Using the 11-point interpolated average precision graph is very informative which is created using 11 cutoff values from 0.0 to 1.0. For each recall level, the arithmetic mean of the interpolated precision at that recall level is calculated for information on each test collection. This graph is one of the most commonly used methods for comparing systems by plotting different runs on the same graph to determine their superiority and performance. The comparisons are suggested to be made at three different recall ranges: low (0.0–0.2), middle (0.2–0.8), and high (0.8–1.0).

4.6. Other measures

Many other evaluation measures have been proposed in the CBIR literature. Manjunath et al. [267] have proposed a performance measure, the averaged normalized modified retrieval rank (ANMRR), which has been used in all MPEG-7 color core experiments. The ANMRR is the evaluation measure in the range [0–1], and the smaller value indicates better matching quality of a query. The calculations of ANMRR are based on the average rank AVR(q) for a given query q, where the Rank(k) of the kth image in the retrieved images is defined as the position at which this image is retrieved. The general form of ANMRR criterion is defined as [267]:

$$\text{ANMRR} = \frac{1}{N_Q}\sum_{q=1}^{N_Q} NMRR_q, \tag{19}$$

where NMRR is a normalized modified retrieval rank that ranges between 0 and 1. Recently, Chatzichristofis et al. [283] have proposed a new performance measure for CBIR systems, referred as the mean normalized retrieval order (MNRO). This measure evaluates retrieval systems by considering the position where each image appears in the ranked list. Unlike the NMRR, the MNRO differs with respect to the parameter which considers an image as non-retrieved if retrieved after position k. In addition, an upper limit is used in the retrieval process that dynamically designated for each query by taking into account the query generality. Therefore, the retrieved images after that limit will still contribute to the performance measure but at lower degree. Using this approach, the new performance measure can predict the behavior of scaled-up system version and not biased on the top-10 or top-20 results.

Petrakis et al. [284] utilized the notion of ranking quality (R_{norm}) which measures the differences between the rankings of results obtained by the retrieval method and human referee. The higher the value of R_{norm} the better the ranking quality of the method, i.e. the method retrieves the qualifying (similar) entries before the non-qualifying ones. The calculations of R_{norm} contain four steps: (1) the answers of the candidate method are evaluated and then judged either qualifying or not qualifying; (2) each answer is assigned a 'rank' which equals its order in the answer set (i.e. the first = rank 1, the second = rank 2 and so on); (3) these ranked answers are formulated in pairs such that only the pairs with one qualifying and one non-qualifying answer are taken. In each pair, the qualifying entry is first and the non-qualifying entry is second; and (4) the relative ranks of answers in each pair are examined and the R_{norm} is computed as:

$$R_{norm} = \begin{cases} \frac{1}{2}\left(1 + \frac{S^+ - S^-}{S^+_{\max}}\right) & \text{if } S^+_{max} > 0; \\ 1 & \text{otherwise.} \end{cases} \tag{20}$$

where S^+ is the number of correctly ranked pairs (i.e. the qualifying entry has higher rank), S^- is the number of erroneously ranked pairs (i.e. the method assigned a higher rank to the non-qualifying entry), and $S^+_{\max}$ is the total number of ranked pairs.

5. Research directions and discussion

This section highlights the most important issues in CBIR domain addressed in this study, and discusses several research issues/directions along with our own thoughts and future insights.

5.1. Research issues

5.1.1. Query specifications and structures

The majority of end-users of current search engines still mainly use the traditional text-based instead of CBIR-based techniques (e.g. search-by-example). Despite that some search engines offer image submission instead of text, they still far away from user expectations. However, users usually do not provide the image that exactly reflects their needs since they are more familiar with textual captions. On the other hand, the text-based query is a subjective measure and usually lakes of semantic concepts which impacts the user satisfaction. As a consequence, using the textual and visual contents simultaneously seems more effective and an inevitable option in order to improve the retrieval accuracy. Recently, research efforts have concentrated on using image datasets with tags (e.g. ImageNet) which allow for the use of both visual and textual contents. Since there is no single standard format or structure for the user query, more investigations are necessary on formulating and modeling a meaningful query that improves the user satisfaction. Additionally, there are other important aspects related to the query structure should be considered such as visualization, visual analytics, and GUIs (see Section 2.5.2). More discussion and highlights on these aspects will be provided later in this section.

5.1.2. Encoding of interest points

Local image features (e.g. SIFT, SURF, and LBP) have been utilized extensively over the last few years since they provide more stability and robust against the local geometric distortion. Since humans are usually interested in certain parts of images (e.g. objects), object-based retrieval which uses local descriptors tend to be more effective in satisfying the human interest. However, global features (e.g. color and texture) also provide a robust representation of the whole image. Therefore, many of research efforts in the CBIR domain have combined global and local features to provide a more representative description for the visual contents of the image. In addition, considering only one image feature, either local or global, is not sufficient to achieve high retrieval accuracy in different CBIR tasks, e.g. image classification and retrieval. As a result, it is expected that more studies will be continuously proposed to utilize local descriptors effectively and to combine them with global features to provide more descriptive image features.

However, image representation is a very important factor in large scale image retrieval. For instance, BOW and Fisher Kernel methods have been successfully employed in the field of computer vision, and we believe that they will be utilized more for different CBIR tasks. In the context of high-dimensional vectors, the image representation/signature is closely related to the methods of dimensionality reduction and indexing which have a crucial impact on the CBIR system performance. There is a trade-off between the dimensionality reduction and the indexing algorithm.

Specifically, high-dimensional image representations usually provide better search results, but it is more difficult to be indexed efficiently. Therefore, it is important to provide a joint optimization of the following constraints: search accuracy, efficiency, and memory usage [285]. The last two factors are related and become critical while considering a dataset with thousands or millions of images. More research efforts are expected to develop an optimized image representation by a moderate dimensional, and the compact code of aggregated image descriptors is one of the effective advances that has been recently proposed by many studies [185,285–287]. These methods optimize the dimensionality reduction and indexing of images while preserving high retrieval accuracy. Precisely, the image is represented into a reduced dimension of lower bytes as a compact representation to obtain efficient vector comparisons and achieve better performance. Such compact codes are mainly needed in the real-world systems, e.g. web search and mobile applications.

5.1.3. Very deep structures

Machine learning has recently become a vital approach associated with information retrieval methods. Deep learning has been successfully employed in several CBIR tasks such as image classification, object recognition and retrieval. It is expected to gain more attention and investigation in the near future. Using such deep structure benefits from its independence from the domain knowledge which allows systems to learn image features and process the data in multiple stages and transformations. Deep learning has proved its considerable improvement when employed for learning feature representations/extraction, distance metrics, and automatic annotation. The CNN is the most deep neural networks that recently studied and employed in computer vision and retrieval. Since these deep structures have mainly utilized for particular tasks (e.g. classification and recognition), it is important to examine the ability of reusing the trained systems for other CBIR tasks but in different domains. Recently, many works [20,288,289] have investigated the extracted image representations by deep CNNs and evaluated the capability of learning and generalizing the trained deep model for other CBIR tasks.

5.1.4. Class-independent similarities

Distance metric learning needs further studies to improve the performance and accuracy of CBIR systems. The main reason is that the human usually looks for semantically related objects in images rather than visually similar images. Accordingly, some new methods for similarity learning have been proposed such as the online algorithm for scalable image similarity learning (OASIS) [290]. In such approaches, the similarity information is extracted from pairs of images that share a common label or are retrieved in response to a common text query. Additionally, it is designed to learn the class-independent similarity measure with no need for class labels over the sparse representation. Despite faster and accurate results obtained by query independent similarity methods, its scalability becomes critical when large datasets considered, e.g. over the web. Moreover, most of the proposed methods for similarity learning are class-dependant retrieval approaches and use the distance measure as a metric. This limits the ability of CBIR systems in retrieving more relevant images that are semantically similar in contents but belong to different image classes. Consequently, it is important to have class-independent similarity measures that able to learn the similarity between all images regardless of their concept category, and to rely on the self-discrimination rather than on the prior knowledge of image content and domain.

5.1.5. Group sparsity and event-based annotation

Automatic image annotation is another successful achievement of using the deep learning approach and expected to be effectively utilized in many CBIR tasks. For web retrieval, generating automatic labels of the given images by deep CNNs is a big advancement and considered as a breakthrough in the field of CBIR. The image description generated by the deep CNN models has a superior of containing some verbs to describe both objects and actions included in the image (see Section 2.4.3). The event-based automatic annotation is beneficial for (1) reducing the wasted effort spent in manual image annotation and tagging, (2) avoiding inaccuracies, (3) improving the discriminating power of retrieval system, (4) providing more semantics for the object-based search and classification, and (5) offering the perfect alternative for users with disabilities to get image descriptions (i.e. captions) by the system. This definitely will inspire more efforts to utilize such approaches for query restructuring in order to simplify the search mechanisms for end-users.

As an integral emerging approach, the sparsity prior has been recently used to solve the challenges associated with image annotation. Basically, the group sparsity mechanism focuses on the features extracted rather than the model representations of keywords. However, these features do not contribute equally or positively to the annotation performance. Specifically, the predefined features may have (1) a sparse prior and can be pruned or assigned different weights, and (2) a group clustering trend, i.e. the nonzero weights exist in the union of subspaces [291]. A considerable works have been introduced to solve the feature-related problems in the image annotation and retrieval applications. Some recent works use feature selection by structural group sparsity for image annotation [291,292], image annotation and classification by multilayer group sparse coding [293], and joint group sparsity for tag localization [294]. In addition to the group sparsity, the produced annotations are beneficial for content-based action recognition (e.g. human actions). For instance, still images based human action recognition [295] has witnessed a considerable attention as a promising approach which focuses on identifying the person's behavior from single image rather than from the video information. These advances are expected to be active research topics in computer vision and information retrieval for a long time.

5.1.6. Descriptive datasets

Humans have rich semantic vocabularies that are difficult to be counted and categorized in a certain manner. Many research works and image datasets are based on a large lexical database of English, namely the WordNet [296], where the nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms which express a distinct concept. Despite the availability of some challenging image datasets such as ImageNet, the existing datasets are not descriptive enough due to (1) the lack of many concepts and (2) the limitation of using only nouns in image captions and tags. This absolutely limits the level of semantic description provided for image objects and actions. Improving the existing image datasets by including more semantic descriptions and using them with image features themselves would be beneficial to improve the semantic-based image retrieval. On the other hand, the use of such image datasets for evaluation and benchmarking is not an easy task. As a result, the ideal image dataset should have some requirements [297]: (1) the test set should be representative for the interesting image retrieval area and cover the entire spectrum of imagery sources; (2) the ground truth should be available for the test set so that objective evaluations can be performed; (3) the test set should be easily accessible and freely redistributable for both researchers and reviewers; and (4) providing a set of standardized tests associated with the dataset in order to perform a comparative benchmarking since different researchers usually perform different performance tests on the same database.

5.1.7. Rank learning

Generally, CBIR systems rank the returned images according to their similarity to a given query image. Because the result set is usually large, users will only inspect the topmost results so their perception about the system quality mainly depends on the results relevance [298]. The typical ranking approaches are based on the extracted image descriptors and their similarities thus different descriptors produce different rankings. It is obvious that the integration of multiple descriptors is better than using a single descriptor and may improve ranking performance, but forming an optimal integration is a data-dependent process which is difficult to obtain in advance. Therefore, it would be better to train the CBIR system on a rank learning scheme to be used for similarity matching. Many studies have been proposed in this context [298–300]. For instance, Faria et al. [298] have proposed a rank learning method which considers the image relevancy to a given query image as input. This information is used for training so that learning algorithms produce a ranking function which maps the similarities to the relevance levels between dataset images and query images. The relevance of images returned for a new query image is estimated according to the learned function which associates a score with each image indicating its relevance to the query image.

In the image retrieval process, rank learning methods have been recently exploited and associated with relevance feedback approaches to minimize the problems of supervised approaches. This trend is mainly based on some unsupervised re-ranking strategies that require no user intervention or labeled data. The key idea is to consider the intrinsic relationships between all dataset objects as post-processing of the similarity scores, and then judge the similarity between objects by considering a specific contexts as humans do [301]. In the re-ranking context, research efforts attempt to improve three performance attributes simultaneously: (1) effectiveness (i.e. the quality of retrieved images), (2) scalability (i.e. the ability of handling growing image dataset), and (3) efficiency (i.e. the retrieval speed). Different re-ranking strategies have been adopted in real-world systems, including web search engines (e.g. Google, Yahoo, and Bing), but the user unsatisfaction on the initial returned results still a big challenge. This inspires the research community in computer vision and multimedia information retrieval to propose many approaches which may improve the ranking of retrieved list. Recently, a comprehensive study [302] has discussed many re-ranking approaches in multimedia retrieval and broadly categorized them as: (1) self-reranking that only based on the initial search results; (2) example-based re-ranking that utilizes the user-provided query examples; (3) crowd re-ranking that discovers the crowdsourcing knowledge available on the internet; and (4) the interactive re-ranking which is based on the user guidance in the re-ranking process.

5.1.8. User-centered interactive design

Users incorporation into the evaluation of information retrieval systems as well as the study of user behaviors and interactions during the search process has a crucial impact on the performance of CBIR systems. As a result, appropriate approaches of studying the interactive information retrieval (IIR) [303,304] systems should pointedly investigate the user interactions with systems and information. Different CBIR systems, GUIs, and search scenarios require different methods, measures, and interaction manners with users. Consequently, user-centered CBIR designs that go beyond the traditional methods for CBIR designs are required. However, IIR studies offer many choices for the developers of CBIR systems about how to design and conduct evaluations, but it lacks enough information about how doing this. Therefore, some related fields along with IIR aspects are important in the user-centered design, including visual analytics and visualization (see Section 2.5.2), psychology, and human computer interaction (HCI). Fig. 9 demonstrates the process of information visualization and interactions between humans and machines as an integrating discipline. These key factors should adapt to the characteristics of CBIR design and support the information workflow on all levels of retrieval operations. As a result, we believe that CBIR developers should consider these two questions: at machine side, does this system retrieve relevant images? And at human side, can users use this system to retrieve relevant images?

5.1.9. Hardware accelerators and parallel computing

The rapid advances in digital and multimedia contents have dual effects in the CBIR domain. On one hand, modern software and hardware largely provide more alternatives for modeling retrieval systems with minimal human interaction and maximal satisfaction. On the other hand, the growing abundance of multimedia content and the associated metadata make the retrieval process more complex, especially for images with different situations, equipment sources, formats, sizes, resolutions, dimensions, and semantic contents. As a result, large-scale image retrieval applications are computationally expensive. Currently, there are many technologies may be utilized to avoid the performance degradation, including high performance computing (HPC), hardware accelerators, and big data. The HPC with a parallel fashion on multiple and cloud resources has the capacity to handle and analyze massive amounts of data at high speed. Some CBIR tasks, e.g. classifier training on huge image datasets which may take weeks/days using the standard computers/processors can be alternatively done

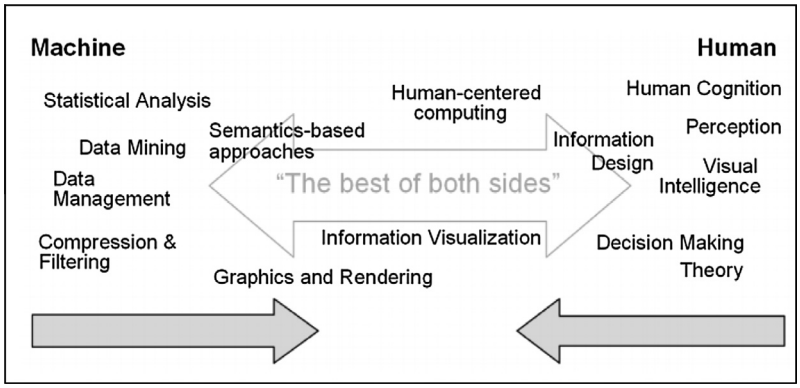

Fig. 9. The key disciplines of machine–human interactions [242].

in a few hours or minutes. Hardware accelerators such as the graphics processing units (GPU) and the field-programmable gate array (FPGA) increasingly become popular resources to assist retrieval systems in performing complex and intensive computations.

GPUs and FPGAs together with other accelerators such as the digital signal processors (DSP) and media/network processors can process work offloaded by CPUs and send the results back upon completion. Generally, utilizing such hardware accelerators will capture and process data within a tolerable elapsed time, especially in complex CBIR tasks such as deep learning and image annotation. The vast parallel computing resources and programming environments of accelerators make them optimal to accelerate the execution of parallel parts in CBIR applications.

5.1.10. Adaptation of domain-specific CBIR applications

General CBIR systems are widely employed in many specific domains such as medical care, object tracking, biometric detections, etc. Images in such domains have particular characteristics that need to be considered while designing CBIR technologies. In medical imaging, millions of medical images (e.g. magnetic resonance imaging (MRIs), X-rays, computed tomography (CT), and 2D modalities) are produced daily worldwide, and the access to these types of images become more complicated. Since the medical imaging is one of the most integral domains that use CBIR applications, it is essential to consider the special needs of such multimodality and multidimensional images and so to adapt them with special medical equipment. This will definitely provide many clinical benefits, improved diagnostics, and more accurate decisions making. Muller et al. [305] and Kumar [306] provide reviews of CBIR systems in the medical domain. Other domains use CBIR approaches to identify humans based on their unique physical traits, e.g. fingerprints, faces, and eyes. This is vital for many applications such as biometrics security, facial and Iris recognition. Moreover, CBIR technologies witness a remarkable interest in the education and learning application that involves huge materials with embedded images which need to be efficiently indexed, retrieved and visualized.

5.1.11. CBIR systems: from laboratories to markets

The majority of proposed CBIR systems are developed as research prototypes being developed in research laboratories. These range from the very early CBIR systems [2–7] (e.g. QBIC, VisualSeek, Photobook and Blobworld) to thousands of prototypes and research projects currently available in the field. The existing CBIR systems offer a wide range of search capabilities such as uploading images, web link (i.e. the URL of image), or drag-and-drop image mechanisms. Additionally, they provide multi-functionality and attractive GUIs which enable users to easily make image selections, relevance feedback, refinement, and history preferences. This also includes an automatic ranking of the retrieved images in an organized list with some functions enabled such as image zooming, saving, and resizing. Some of these CBIR systems become available nowadays for the public as standalone or part of commercial products. Several web search engines employ the CBIR service to people. Google image search, TinyEye, and Querbie are some examples of web search engines that support search-by-example capabilities for large-scale image retrieval; i.e. millions of images.

Other systems such as Apple's built-in iPhoto application in Mac OS and Pikalike offer many editing services for images, e.g. face recognition, similarity based sorting, and automatic tagging for image contents. Such these applications are also offered as a standalone application for particular machines (e.g. smart phones and tablets). Additionally, other CBIR functionalities are provided for E-commerce applications over the web such as eBay's fashion and Amazon. LabelMe [307] and ALIPR [308] have been also developed for automatic image annotation and tagging. Along with the existing CBIR systems currently available, a considerable amount of systems are expected to be developed in order to meet the rapidly growing needs of users, especially for image services in social media, education, and medical care. However, the availability of CBIR systems to the public, including researchers and end-users, is restricted by the licensing and use agreements which manage all what people can do with such systems. However, some of CBIR systems are freely available to the public under the general public license (GPL) as open source systems or as application program interfaces (APIs), while other systems are closed and license-based applications for commercial purposes.

5.2. Revisiting the CBIR framework

This section summarizes and highlights the most important issues addressed in this study along with the questions raised in Section 1. Table 9 lists issues/blocks of the CBIR framework (see Section 2) against some relevant works/methods/papers addressed

Table 9
Main issues addressed in the CBIR framework.

Main issues/challenges/mechanisms	Related works
Features combination	Pseudo-Zernike moments [107], Gabor filters [128], Wavelets [133], MSD [129], FCTH [130], color-texture-shape [136], LOCTP [168], color-texture [172]
Optimization approaches used as part of other methods	Color Correlogram [105], Color moments [104], DCD [114], DBPSP [172], PCA [183], BOW [184], FV [185]
Robust feature representations against image transformations	Pseudo-Zernike moments [107], DCV [117], FCTH [130], Shape Google [140], Rolling Penetrate [143], SIFT [158], SURF [161], LBP [161]
Compact feature representations	BTC [111], DCD [114], PCA-SIFT [158], GIST [178], VLAD [187], Compact codes [285–287]
The use of spatial information	CCM [109], Gabor filters [118,146–151,120,152–157]
The problem of 'Curse of dimensionality'	GIST [178], KPCA [183], BOW [184], Fisher Kernels [185], VLAD [187], FLDA [188], Manifold learning [190–193], Tree-based indexing [194–197], Hash-based indexing [198–204], LSI [206,207]
The use of machine learning	Unsupervised learning [210–212], Supervised learning [219–226], Deep learning [20,227–232], Distance metric learning [261–266], Rank learning [299–303]
Automatic image tagging/annotation	CNN-based [231], Group sparsity [291–295]
Relevance feedback and refinement	RF in CBIR [18,233–241], VA [242–246]
Human-system interactions	VA [242–246], Visualization [247–252]
Similarity/Dissimilarity	BTC [111], Histogram-based measures [253,255], EMD [254], Graph-based [256–260]
Performance evaluation	ANMRR [267], MNRO [283], ranking quality [284]
Benchmarking and evaluation challenges	Pascal-Voc [280], ImageCLEF [281], ILSVR [282], and standard measures
CBIR domains and applications	Medical imaging [306,307], Google search-by-example, Apple's iPhoto, eBay fashion. LabelMe [308], ALIPR [312], Early CBIR systems [2–7]

them. Finally, all of research trends and emerging technologies in the field of CBIR should be organized in a certain robust architecture to provide more constructive impacts on the performance of real-world systems such as: web search, social media retrieval, mobile applications, medical purposes, surveillance, geolocation, and facial recognition.

6. Conclusions

This paper has presented a comprehensive survey on different techniques and recent research works in the CBIR domain. Firstly, this study has discussed the general retrieval framework which most of CBIR systems have adopted over the last 15 years. Secondly, the proposed approaches in each block of CBIR framework have been presented along with the impact of research shift from low-level to high-level processing. Thirdly, this study has identified the most recent advances that contribute in reducing the semantic gap. Finally, it has identified some important issues that directly affect CBIR systems as follows: (1) image representation with focus on local descriptors; (2) automatic image annotation which opens the research gate for action-based image retrieval; (3) dimensionality reduction and image indexing that directly affect the CBIR performance in terms of accuracy, speed, and memory usage; (4) deep learning as a promising approach for reducing the semantic gap, especially the use of deep CNNs for classification, distance metric learning, relevance feedback, and annotation; (5) a description of ideal image datasets used for training and testing in the CBIR context; (6) re-ranking approaches associated with relevance feedback as post-processing to minimize users intervention and satisfy their needs; and (7) visualization aspects and the importance of modeling CBIR systems that semantically handle and return the most relevant images with maximum user satisfaction. Additionally, this study has presented some recent interesting and promising techniques that are expected to make a breakthrough in the field of image retrieval. These

Table 10
Definitions of all acronyms used throughout this paper.

Acronym	Definition	Acronym	Definition	Acronym	Definition
AAM	Active Appearance Model	**EC**	Evolutionary Computation	**LSI**	Latent Semantic Indexing
AC	Active Contours	**EHD**	Edge Histogram Descriptor	**MCM**	Motif Co-Occurrence Matrix
ALBNN	Algorithm Learning Based Neural Network	**EMD**	Earth Mover's Distance	**MST**	Minimal Spanning Tree
ANMRR	Averaged Normalized Modified Retrieval Rank	**ESS**	Efficient Subwindow Search	**MRF**	Markov Random Field
ANN	Artificial Neural Networks	**EM**	Expectation Maximization	**MSD**	Micro- Structure Descriptor
API	Application Program Interface	**FPGA**	Field-Programmable Gate Array	**MTH**	Multi-Texton Histogram
ASM	Active Shape Model	**FCTH**	Fuzzy Color And Texture Histogram	**MPCK**	Metric Pairwise Constrained K-means
BCCM	Block Color Co-Occurrence Matrix	**FLDA**	Fisher Linear Discriminant Analysis	**MNRO**	Mean Normalized Retrieval Order
BPH	Block Pattern Histogram	**FSVM**	Fuzzy Support Vector Machine	**MRI**	Magnetic Resonance Imaging
BTC	Block Truncation Coding	**GBDA**	Generalized BDA	**NPRF**	Navigation-Pattern-Based Relevance Feedback
BOW	Bag-Of-Words	**GMM**	Gaussian Mixture Models	**OLS**	Ordinary Least Squares
BDEE	Biased Discriminative Euclidean Embedding	**GUI**	Graphical User Interface	**ORB**	Oriented Fast And Rotated Brief
BMMA	Biased Maximum Margin Analysis	**GMRF**	Gaussian Markov Random Field	**OASIS**	Online Algorithm For Scalable Image Similarity
BDA	Biased Discriminant Analysis	**GLCM**	Gray-Level Co-Occurrence Matrix	**PCA**	Principal Components Analysis
BFS	Breadth-First Search	**GLOH**	Gradient Location-Orientation Histogram	**PDE**	Partial Differential Equation
CBR	Case-Based Reasoning	**GPU**	Graphics Processing Units	**QR**	Query Reweighting
CBIR	Content –Based Image Retrieval	**GPL**	General Public License	**QPM**	Query Point Movement
CCM	Color Co-Occurrence Matrix	**HCI**	Human Computer Interaction (HCI	**QEX**	Query Expansion
CCV	Color Coherence Vector	**HPC**	High Performance Computing	**RAG**	Region Adjacency Graph
CSS	Curvature Scale Space	**IIR**	Interactive Information Retrieval	**ROI**	Region Of Interests
CPDH	Contour Points Distribution Histogram	**IR**	Information Retrieval	**RR**	Ridge Regression
CT	Curvature Tree-Based	**ISOMAP**	Isometric Mapping	**RNN**	Recurrent Neural Networks
CLBP	Completed LBP	**ICA**	Independent Component Analysis	**SPD**	Steerable Pyramid Decomposition
CLBPC	CLBP-Center	**IGA**	Interactive Genetic Algorithm	**SED**	Structure Element Descriptor
CLBPS	CLBP-Sign	**IS**	Information Seeking	**SHE**	The Structure Element Histogram
CLBPM	CLBP-Magnitude	**KPCA**	Kernel-PCA	**SIFT**	Scale Invariant Feature Transform
CHoG	Compressed Histogram Of Gradients	**KNN**	K Nearest Neighbors	**SURF**	Speeded Up Robust Features
CNN	Convolutional Neural Networks	**LASSO**	Least Absolute Shrinkage And Selection Operator	**SVD**	Singular Value Decomposition
CT	Computed Tomography	**LMS**	Least Mean Square	**SSH**	Sparse Spectral Hashing
DAN2	Dynamic Artificial Neural Network	**LBP**	Local Binary Patterns	**SSM**	Statistical Shape Models
DWT	Discrete Wavelet Transform	**LBPV**	LBP Variance	**SUI**	Search User Interface
DCD	Dominant Color Descriptor	**LDSMT**	Local Difference Sign-Magnitude Transforms	**TF-IDF**	Term Frequency Inverse Document Frequency
DCV	Distance Coherence Vector	**LDP**	Local Derivative Patterns	**TML**	Transfer Metric Learning
DCT	Discrete Cosine Transform	**LTP**	Local Ternary Pattern	**VAMIR**	Visual Analytics For Medical Image Retrieval
DOG	Difference-Of-Gaussian	**LTrPs**	Local Tetra Patterns	**VAR**	Rotation Invariant Variance
DLBP	Dominant-LBP	**LOCTP**	Local Oppugnant Color Texture Pattern	**VLAD**	Vector Of Locally Aggregated Descriptors
DBPSP	Difference Between Pixels Of Scan Pattern	**LFDA**	Local Fisher Discriminant Analysis	**VOM2FNN**	Vectorization-Optimization-Method-Based Type-2 Fuzzy Neural Network
DeepID	Deep Hidden Identity Features	**LLE**	Locally Linear Embedding	**VOM**	Vectorization-Optimization Method
DSDC	Differential Scatter Discriminant Criterion	**LPP**	Locality Preserving Projections		
DSP	Digital Signal Processors	**LSH**	Local Sensitivity Hashing		

techniques should vastly contribute in solving the challenges in CBIR domain such as the: semantic gap, curse of dimensionality, comprehensive image representation, automatic image annotation, ranking, and visual structure of the query. Image datasets, distance measures, and performance measures are also very important components that need to be more standardized in order to provide a robust ground for CBIR evaluations in terms of accuracy, efficiency, and scalability. Consequently, a comprehensive retrieval model, which is not limited to some components and tasks in the CBIR domain, becomes an exigent demand. A successful employment of any proposed CBIR system in a certain task (e.g. object recognition) is expected to be generalized for other CBIR tasks (e.g. automatic image annotation). Finally, designing a user-centered CBIR system can be considered as an integrating discipline, i.e. application and domain specific research areas should contribute with their existing procedures and models. Such smart and semantic-based CBIR systems will extremely contribute in many real-world systems such as web search, social media, mobile, and medical applications.

Appendix A

See Table 10.

References

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.

[2] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, G. Taubin, The QBIC project: Querying images by content using color, texture and shape, in: Proceedings of the SPIE Storage and Retrieval for Image and Video Databases, San Jose, CA, 1994.

[3] J.R. Smith, S.F. Chang, VisualSEEk: a fully automated content-based image query system, in: Proceedings of the Forth ACM International Conference on Multimedia '96, Boston, MA, 1996.

[4] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 947–963.

[5] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1026–1038.

[6] J.R. Smith, S.F. Chang, Visually searching the Web for content, IEEE Multim. 4 (1997) 12–20.

[7] S. Sclaroff, M. LaCascia, S. Sethi, L. Taycher, Unifying textual and visual cues for content-based image retrieval on the World Wide Web, Comp. Vis. Image Understand. 75 (1999) 86–98.

[8] X.S. Zhou, T.S. Huang, CBIR: from low-level features to high level semantics, in: Proceedings of the SPIE, Image and Video Communication and Processing, vol. 3974, San Jose, CA, 2000, pp. 426–431.

[9] R. Brunelli, O. Mich, Image retrieval by examples, IEEE Trans. Multim. 2 (3) (2000). p. 164, 171.

[10] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. ISBN0-387-31073-8.

[11] RC. Veltkamp, M. Tanase, Content-Based Image Retrieval Systems: A Survey, rapport no UU-CS-2000-34, 2000.

[12] C. Jörgensen, Image Retrieval: Theory and Research, Scarecrow Press, 2003.

[13] R. Datta, J. Li, JZ. Wang, Content-based image retrieval: approaches and trends of the new age, in: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2005, pp. 253–262.

[14] M.S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: state of the art and challenges, ACM Trans. Multim. Comput., Commun. Appl. 2 (1) (2006) 1–19.

[15] Y. Liu, D. Zhang, G. Lu, WY. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recog. 40 (1) (2007) 262–282.

[16] R. Datta, D. Joshi, J. Li, J. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. (CSUR) 40 (2) (2008) 5.

[17] R. Priyatharshini, S. Chitrakala, Association based image retrieval: a survey, in: Mobile Communication and Power Engineering, Springer, Berlin Heidelberg, 2013, pp. 17–26.

[18] J. Li, N.M. Allinson, Relevance feedback in content-based image retrieval: a survey, in: Handbook on Neural Information Processing, Springer, Berlin Heidelberg, 2013, pp. 433–469.

[19] L. Ai, J. Yu, Y. He, T. Guan, High-dimensional indexing technologies for large scale content-based image retrieval: a review, J. Zhejiang Univ. Sci. C 14 (7) (2013) 505–520.

[20] J. Wan, D. Wang, SCH. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: a comprehensive study, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 157–166.

[21] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural Codes for Image Retrieval, 2014 Available from arXiv:1404.1777.

[22] B. Verma, S. Kulkarni, Neural networks for content based image retrieval, Seman.-Based Vis. Inf. Ret. (2006) 252–272.

[23] R.C. Gonzalez, R.E. Woods, Digital image processing, 2002.

[24] F. Jing, M. Li, H.J. Zhang, B. Zhang, An efficient and effective region-based image retrieval framework, IEEE Trans. Image Process. 13 (5) (2004) 699–709.

[25] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comp. Vis. 59 (2) (2004) 167–181.

[26] L. Grady, Multilabel random walker segmentation using prior models, in: IEEE Conference of Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 763–770.

[27] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.

[28] L. Luccheseyz, S.K. Mitray, Color image segmentation: a state-of-the-art survey, Proc. Ind. Nat. Sci. Acad. (INSA-A) 67 (2) (2001) 207–221.

[29] H. Zhang, J.E. Fritts, S.A. Goldman, Image segmentation evaluation: a survey of unsupervised methods, Comp. Vis. Image Understand. 110 (2) (2008) 260–280.

[30] C. Jung, J. Liu, T. Sun, L. Jiao, Y. Shen, Automatic image segmentation using constraint learning and propagation, Dig. Sig. Process. 24 (2014) 106–116.

[31] F.J. Estrada, A.D. Jepson, Benchmarking image segmentation algorithms, Int. J. Comp. Vis. 85 (2) (2009) 167–181.

[32] F. Wang, Q. Huang, M. Ovsjanikov, LJ. Guibas, Unsupervised multi-class joint image segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3142–3149.

[33] B. Peng, L. Zhang, D. Zhang, A survey of graph theoretical approaches to image segmentation, Pattern Recog. 46 (3) (2013) 1020–1038.

[34] N. Senthilkumaran, R. Rajesh, Edge detection techniques for image segmentation–a survey of soft computing approaches, Int. J. Rec. Trends Eng. 1 (2) (2009).

[35] Z. Wu, R. Leahy, An optimal graph theoretic approach to data clustering: theory and its application to image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 15 (11) (1993) 1101–1113.

[36] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[37] C. Zhong, M. Malinen, D. Miao, P. Fränti, A fast minimum spanning tree algorithm based on k-means, Inf. Sci. 295 (2014) 1–17.

[38] H.N. Gabow, Z. Galil, T.H. Spencer, R.E. Tarjan, Efficient algorithms for finding minimum spanning trees in undirected and directed graphs, Combinatorica 6 (1986) 109–122.

[39] W.B. March, P. Ram, A.G. Gray, Fast Euclidean minimum spanning tree: algorithm analysis and applications, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010.

[40] R.G. Gallager, P.A. Humblet, P.M. Spira, A distributed algorithm for minimum-weight spanning trees, ACM Trans. Program. Lang. Syst. 5 (1983) 66–77.

[41] X. Wang, X. Wang, D.M. Wilkes, A divide-and-conquer approach for minimum spanning tree-based clustering, IEEE Trans. Knowl. Data Eng. 21 (2009) 945–958.

[42] C. Lai, T. Rafa, D.E. Nelson, Approximate minimum spanning tree clustering in high-dimensional space, Intell. Data Anal. 13 (2009) 575–597.

[43] C. Zhong, M. Malinen, D. Miao, P. Fränti, A fast minimum spanning tree algorithm based on k-means, Inf. Sci. (2014).

[44] S.H. Kwok, A.G. Constantinides, A fast recursive shortest spanning tree for image segmentation and edge detection, IEEE Trans. Image Process. 6 (2) (1997) 328–332.

[45] M. Stoer, F. Wagner, A simple min-cut algorithm, J. ACM (JACM) 44 (4) (1997) 585–591.

[46] D. Nanongkai, H.H. Su, Almost-tight distributed minimum cut algorithms, in: Distributed Computing, Springer, Berlin Heidelberg, 2014, pp. 439–453.

[47] B. Ghanem, N. Ahuja, Dinkelbach, NCUT: an efficient framework for solving normalized cuts problems with priors and convex constraints, Int. J. Comput. Vis. 89 (1) (2010) 40–55.

[48] D.S. Hochbaum, Polynomial time algorithms for ratio regions and a variant of normalized cut, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 32 (5) (2010) 889–898.

[49] A. Fabijanska, Normalized cuts and watersheds for image segmentation, in: IET Conference Publications (600 CP), 2012.

[50] A. Sáez, C. Serrano, B. Acha, Normalized cut optimization based on color perception findings: a comparative study, Mach. Vis. Appl. 25 (7) (2014) 1813–1823.

[51] P. Kohli, PHS Torr, Efficiently solving dynamic markov random fields using graph cuts, in: ICCV 2005, Tenth IEEE International Conference on Computer Vision, vol. 2, 2005, pp. 922–929.

[52] Y. Boykov, G. Funka-Lea, Graph cuts and efficient N-D image segmentation, Int. J. Comp. Vis. 70 (2) (2006) 109–131.

[53] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[54] D. Freedman, T. Zhang, Interactive graph cut based segmentation with shape priors, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, 2005, pp. 755–762.

[55] O. Veksler, Star shape prior for graph-cut image segmentation, in: European Conference on Computer Vision, 2008, pp. 454–467.
[56] V. Lempitsky, P. Kohli, C. Rother, T. Sharp, Image segmentation with a bounding box prior, in: IEEE International Conference on Computer Vision, 2009, pp. 277–284.
[57] J. Liu, J. Sun, HY. Shum, Paint selection, ACM Trans. Graph. (ToG) 28(3) (2009) 69.
[58] A.X. Falcao, J.K. Udupa, F.K. Miyazawa, An ultra-fast user-steered image segmentation paradigm: live wire on the fly, IEEE Trans. Med. Imag. 19 (1) (2000) 55–62.
[59] X. Bai, G. Sapiro, A geodesic framework for fast interactive image and video segmentation and matting, in: IEEE 11th International Conference on Computer Vision. ICCV (2007), pp. 1–8.
[60] P.F. Felzenszwalb, D.P. Huttenlocher, Image segmentation using local variation, in IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 98–104.
[61] Y. Weiss, Segmentation using eigenvectors: a unifying view, In Computer vision, the proceedings of the seventh IEEE international conference on, vol. 2 (1999), pp. 975-982.
[62] S. Sarkar, K.L. Boyer, Quantitative measures of change based on feature organization: eigenvalues and eigenvectors, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1996.
[63] L. Grady, G. Funka-Lea, Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials, in: Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, Springer, Berlin Heidelberg, 2004, pp. 230–245.
[64] L. Grady, Random walks for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1768–1783.
[65] R. Shen, I. Cheng, J. Shi, A. Basu, Generalized random walks for fusion of multi-exposure images, IEEE Trans. Image Process. 20 (12) (2011) 3634–3646.
[66] M. Pavan, M. Pelillo, A new graph-theoretic approach to clustering and segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, p. I-145.
[67] M. Pavan, M. Pelillo, Dominant sets and pairwise clustering, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 167–172.
[68] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Addison Wesley, Reading, MA, 1992.
[69] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (1986) 679–698.
[70] B. Paul, L. Zhang, X. Wu, Canny edge detection enhancement by scale multiplication, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1485–1490.
[71] Y.T. Hsiao, C.L. Chuang, J.A. Jiang, C.C. Chien, A contour based image segmentation algorithm using morphological edge detection, in: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2005, pp. 2962–2967.
[72] G. Sapiro, Vector (self) snakes: a geometric framework for color, texture, and multiscale image segmentation, in: Proceedings of the International Conference on Image Processing, vol. 1, 1996, pp. 817–820.
[73] X. Yu, J. Yla-Jaaski, A new algorithm for image segmentation based on region growing and edge detection, in: Proceedings of the IEEE International Symposium on Circuits and Systems, 1991, pp. 516–519.
[74] H.G. Kaganami, Z. Beij, Region based detection versus edge detection, IEEE Trans. Intell. Inf. Hid. Multim. Signal Process. (2009) 1217–1221.
[75] H. Samet, The quadtree and related hierarchical data structures, ACM Comput. Surv. (CSUR) 16 (2) (1984) 187–260.
[76] T. Pavlidis, Structural Pattern Recognition, Springer, New York, 1980.
[77] D.K. Panjwani, G. Healey, Markov random field models for unsupervised segmentation of textured color images, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-17 (10) (1995) 939–954.
[78] A. Trémeau, P. Colantoni, Regions adjacency graph applied to color image segmentation, IEEE Trans. Image Process. 9 (4) (2000) 735–744.
[79] L. Vincent, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Trans. Pattern Anal. Machine Intell. 13 (1991) 499–506.
[80] Y. Zhou, S. Jiang, M. Yin, A region-based image segmentation method with mean-shift clustering algorithm, in: Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'08, vol. 2, 2008, pp. 366–370.
[81] C. Cigla A. Alatan, Region-based image segmentation via graph cuts, in: Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 2272–2275.
[82] I. Karoui, R. Fablet, J. Boucher, J. Augustin, Unsupervised region-based image segmentation using texture statistics and level-set methods, in: Proceedings of the WISP IEEE International Symposium on Intelligent Signal Processing, 2007, pp. 1–5.
[83] B.M. Carvalho, C.J. Gau, G.T. Herman, T.Y. Kong, Algorithms for fuzzy segmentation, Pattern Anal. Appl. 2 (1) (1999) 73–81.
[84] B.M. Carvalho, G.T. Herman, T.Y. Kong, Simultaneous fuzzy segmentation of multiple objects, Discr. Appl. Math. 151 (1) (2005) 55–77.
[85] J. Udupa, P. Saha, R. Lotufo, Relative fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1485–1500.
[86] G.T. Herman, B.M. Carvalho, Multiseeded segmentation using fuzzy connectedness, IEEE Trans. Pattern Anal. Mach. Intell. 23 (5) (2001) 460–474.
[87] Y. Liang, M. Zhang, W.N. Browne, Image segmentation: a survey of methods based on evolutionary computation, in: Simulated Evolution and Learning, Springer International Publishing, 2014, pp. 847–859.
[88] JC. Rubio, J. Serrat, A. López, N. Paragios, Unsupervised co-segmentation through region matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 749–756.
[89] S. Vicente, C. Rother, V. Kolmogorov, Object cosegmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2217–2224.
[90] A. Joulin, F. Bach, J. Ponce, Discriminative clustering for image co-segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1943–1950.
[91] C. Rother, T. Minka, A. Blake, V. Kolmogorov, Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFS, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 993–1000.
[92] M. Sezgin, Survey over image thresholding techniques and quantitative performance evaluation, J. Electron. Imag. 13 (1) (2004) 146–168.
[93] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Understand. 61 (1) (1995) 38–59.
[94] H. Li, O. Chutatape, Automated feature extraction in color retinal images by a model based approach, IEEE Trans. Biomed. Eng. 51 (2) (2004) 246–254.
[95] S. Un, C. Bauer, R. Beichel, Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach, IEEE Trans. Med. Imaging 31 (2) (2012) 449–460.
[96] J. Schmid, J. Kim, N. Magnenat-Thalmann, Robust statistical shape models for MRI bone segmentation in presence of small field of view, Med. Image Anal. 15 (1) (2011) 155–168.
[97] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[98] R. Beichel, H. Bischof, F. Leberl, M. Sonka, Robust active appearance models and their application to medical image analysis, IEEE Trans. Med. Imag. 24 (9) (2005) 1151–1169.
[99] G. J. Edwards, C. J. Taylor, T. F. Cootes, Interpreting face images using active appearance models, Proc. Int. Conf. Face and Gesture Recognition (1998), pp. 300–305.
[100] W. Fang, K.L. Chan, Statistical shape influence in geodesic active contours, in: IEEE Conference on Computer Vision Pattern, vol. 40, 2007, pp. 2163–2172.
[101] S. Ali, A. Madabhushi, An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery, IEEE Trans. Med. Imag. 31 (7) (2012) 1448–1460.
[102] V. Caselles, J.M. Morel, G. Sapiro, A. Tannenbaum, Introduction to the special issue on partial differential equations and geometry-driven diffusion in image processing and analysis (special issue), IEEE Trans. Image Process. 7 (3) (1998) 269–473.
[103] D. Cremers, M. Rousson, R. Deriche, A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape, Int. J. Comp. Vis. 72 (2) (2007) 195–215.
[104] D. Xiaoyin, Image retrieval using color moment invariant, in: The Seventh International Conference on Information Technology: New Generations (ITNG), Las Vegas, NV, 12–14, 2010, pp. 200–203.
[105] H. Jing, S.R. Kumar, M. Mitra, W.J. Zhu, R. Zabih, Image indexing using color correlograms, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 762–768.
[106] G.P. Qiu, Color image indexing using BTC, IEEE Trans. Image Process. 12 (1) (2003) 93–101.
[107] X.Y. Wang, B.B. Zhang, H.Y. Yang, Content-based image retrieval by integrating color and texture features, Multim. Tools Appl. 68 (3) (2014) 545–569.
[108] A. Khotanzad, Y.H. Hong, Invariant image recognition by Zernike moments, IEEE Trans. Pattern Anal. Mach. Intell. 12 (5) (1990) 489–497.
[109] H. Zhang, Z.F. Dong, H. Shu, Object recognition by a complete set of pseudo-Zernike moment invariants, in: 35th IEEE International Conference on Acoustics Speech and Signal Processing, IEEE Press, New York, 2010, pp. 930–933.
[110] V. Kovalev, M. Petrou, Multidimensional co-occurrence matrices for object recognition and matching, Graph. Models Image Process. 58 (3) (1996) 187–197.
[111] N. Jhanwar, S. Chaudhuri, G. Seetharaman, B. Zavidovique, Content based image retrieval using motif co-occurrence matrix, Image Vis. Comput. 22 (14) (2004) 1211–1220.
[112] G. Qiu, Color image indexing using BTC, IEEE Trans. Image Process. 12 (1) (2003) 93–101.
[113] J. Mathews, M.S. Nair, L. Jo, A novel color image coding technique using improved BTC with k-means quad clustering, in: Advances in Signal Processing and Intelligent Recognition Systems, Springer International Publishing, 2014, pp. 347–357.
[114] J.M. Guo, H. Prasetyo, JH. Chen, Content-based image retrieval using error diffusion block truncation coding features, IEEE Trans. Circ. Syst. Video Technol. PP (99) (2014). pp. 1,1.
[115] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, Chichester, 2002.
[116] H. Shao, Y. Wu, W. Cui, J. Zhang, Image retrieval based on MPEG-7 dominant color descriptor, in: The 9th International Conference for Young Computer Scientists, ICYCS, 2008, pp. 753–757.
[117] R. Min, H.D. Cheng, Effective image retrieval using dominant color descriptor and fuzzy support vector machine, Pattern Recog. 42 (1) (2009) 147–157.

[118] J. Zeng, L. Xiupeng, F. Yu, Multiscale distance coherence vector algorithm for content-based image retrieval, Scient. World J. (2014). Article ID 615973, 13 pages.
[119] R. Lukac, B. Smolka, K. Martin, K.N. Plataniotis, A.N. Venetsanopoulos, Vector filtering for color imaging, Sig. Process. Magaz., IEEE 22 (1) (2005) 74–86.
[120] N. Shrivastava, V. Tyagi, Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching, Inf. Sci. 259 (2014) 212–224.
[121] R. Haralick, Statistical and structural approaches to texture, Proc. IEEE 67 (1979) 786–804.
[122] G. Cross, A. Jain, Markov random field texture models, IEEE Trans. Pattern Anal. Mach. Intell. 5 (1) (1983) 25–39.
[123] D.K. Park, Y.S. Jeon, C.S. Won, Efficient use of local edge histogram descriptor, in: Proceedings of the ACM workshops on Multimedia, ACM, 2000, pp. 51–54.
[124] E.P. Simoncelli, W.T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: International Conference on Image Processing, vol. 3, IEEE Computer Society, 1995, p. 3444.
[125] H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, IEEE Trans. Syst., Man Cybernet. 8 (6) (1978) 460–473.
[126] Y. Meyer, Wavelets: Algorithms and Applications, SIAM, Philadelphia, 1993.
[127] G.A. Papakostas, D.E. Koulouriotis, V.D. Tourassis, Feature extraction based on wavelet moments and moment invariants in machine vision systems, Hum.-Cent. Mach. Vis. (2012).
[128] L. Chen, G. Lu, D. Zhang, Effects of different gabor filter parameters on image retrieval by texture, in International Conference on Multi-Media Modeling, IEEE Computer Society, 2004, p. 273.
[129] X. Wang, Z. Wang, A novel method for image retrieval based on structure elements' descriptor, J. Vis. Commun. Image Represent. 24 (1) (2013) 63–74.
[130] G.H. Liu, Z.Y. Li, L. Zhang, Y. Xu, Image retrieval based on micro-structure descriptor, Pattern Recog. 44 (9) (2011) 2123–2133.
[131] S.A. Chatzichristofis, Y.S. Boutalis, Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval, in Ninth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'08, 2008, pp. 191–196.
[132] R. Kwitt, A. Uhl, Lightweight probabilistic texture retrieval, IEEE Trans. Image Process. 19 (1) (2010) 241–253.
[133] N.E. Lasmar, Y. Berthoumieu, Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms, IEEE Trans. Image Process. 23 (5) (2014) 2246–2261.
[134] X.Y. Wang, Z.F. Chen, J.J. Yun, An effective method for color image retrieval based on texture, Comp. Stand. Interf. 34 (1) (2012) 31–35.
[135] C.C. Lai, Y.C. Chen, A user-oriented image retrieval system based on interactive genetic algorithm, IEEE Trans. Instrument. Measur. 60 (10) (2011) 3318–3325.
[136] X.Y. Wang, Y.J. Yu, H.Y. Yang, An effective image retrieval scheme using color, texture and shape features, Comp. Stand. Interf. 33 (1) (2011) 59–68.
[137] J. Vogel, S. Bernt, Performance evaluation and optimization for content-based image retrieval, Pattern Recog. 39 (5) (2006) 897–909.
[138] G.H. Liu, L. Zhang, Y.K. Hou, Z.Y. Li, J.Y. Yang, Image retrieval based on multi-texton histogram, Pattern Recog. 43 (7) (2010) 2380–2389.
[139] B. Julesz, A brief outline of the text on theory of human vision, Trends Neurosci. 7 (2) (1984) 41–45.
[140] A.M. Bronstein, M.M. Bronstein, L.J. Guibas, M. Ovsjanikov, Shape Google: geometric words and expressions for invariant shape retrieval, ACM Trans. Graph. (TOG) 30 (1) (2011). pp. 1.
[141] M. Ovsjanikov, A.M. Bronstein, M.M. BRONSTEIN, L.J. Guibas, Shape Google: A computer vision approach to invariant shape retrieval, in: Proceedings of the Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA'09), 2009.
[142] X. Shu, X.J. Wu, A novel contour descriptor for 2D shape matching and its application to image retrieval, Image Vision Comput. 29 (4) (2011) 286–294.
[143] Y.W. Chen, C.L. Xu, Rolling penetrate descriptor for shape-based image retrieval and object recognition, Pattern Recog. Lett. 30 (9) (2009) 799–804.
[144] S. Wang, D. Liu, F. Gu, H. Feng, L. Yang, Similar matching for images with complex spatial relations, J. Comput. Inf. Syst. 8 (2012) 8727–8734.
[145] T. Jaworska, J. Kacprzyk, N. Marín, S. Zadrożny, On dealing with imprecise information in a content based image retrieval system, in: Computational Intelligence for Knowledge-Based Systems Design, Springer, Berlin Heidelberg, 2010, pp. 149–158.
[146] M.J. Hsiao, Y.P. Huang, T. Tsai, T.W. Chiang, An efficient and flexible matching strategy for content-based image retrieval, Life Sci. J. 7 (1) (2010) 99–106.
[147] B.G. Prasad, K.K. Biswas, S.K. Gupta, Region-based image retrieval using integrated color, shape, and location index, Comp. Vis. Image Understand. 94 (1) (2004) 193–233.
[148] Q. Tian, Y. Wu, T.S. Huang, Combine user defined region-of-interest and spatial layout for image retrieval, in: IEEE International Conference on Image Processing Proceedings, vol. 3, 2000, pp. 746–749.
[149] J. Lee, J. Nang, Content-based image retrieval method using the relative location of multiple ROIs, Adv. Electr. Comp. Eng. 11 (3) (2011) 85–90.
[150] Y.K. Chan, Y.A. Ho, Y.T. Liu, R.C. Chen, A ROI image retrieval method based on CVAAO, Image Vis. Comput. 26 (11) (2008) 1540–1549.
[151] B. Moghaddam, H. Biermann, D. Margaritis, Regions-of-interest and spatial layout for content-based image retrieval, Multim. Tools Appl. 14 (2) (2001) 201–210.
[152] E.G. Petrakis, Design and evaluation of spatial similarity approaches for image retrieval, Image Vis. Comput. 20 (1) (2002) 59–76.
[153] N. Alajlan, M.S. Kamel, G.H. Freeman, Geometry-based image retrieval in binary image databases, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 1003–1013.
[154] N.V. Hoàng, V. Gouet-Brunet, M. Rukoz, M. Manouvrier, Embedding spatial information into image content description for scene retrieval, Pattern Recog. 43 (9) (2010) 3013–3024.
[155] H. Bunke, K. Riesen, Improving vector space embedding of graphs through feature selection algorithms, Pattern Recog. 44 (9) (2011) 1928–1940.
[156] A. Kumar, J. Kim, L. Wen, M. Fulham, D. Feng, A graph-based approach for the retrieval of multi-modality medical images, Med. Image Anal. 18 (2) (2014) 330–342.
[157] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comp. Vis. 60 (2) (2004) 91–110.
[158] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. 2(27), 2004, pp. 506–513.
[159] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[160] H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, Speeded-up robust features (SURF), Comp. Vis. Image Understand. 110 (3) (2008) 346–359.
[161] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[162] S. Liao, M.W.K. Law, A.C.S. Chung, Dominant local binary patterns for texture classification, IEEE Trans. Image Process. 18 (5) (2009) 1107–1118.
[163] Z. Guo, L. Zhang, D. Zhang, Rotation invariant texture classification using LBP variance (LBPV) with global matching, Pattern Recog. 43 (3) (2010) 706–719.
[164] Z. Guo, D. Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Trans. Image Process. 19 (6) (2010) 1657–1663.
[165] B. Zhang, Y. Gao, S. Zhao, J. Liu, Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor, IEEE Trans. Image Process. 19 (2) (2010) 533–544.
[166] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans. Image Process. 19 (6) (2010) 1635–1650.
[167] S. Murala, R.P. Maheshwari, R. Balasubramanian, Local tetra patterns: a new feature descriptor for content-based image retrieval, IEEE Trans. Image Process. 21 (5) (2012) 2874–2886.
[168] I.J. Jeena, K.G. Srinivasagan, K. Jayapriya, Local oppugnant color texture pattern for image retrieval system, Pattern Recog. Lett. 42 (2014) 72–78.
[169] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. 1, 2005, pp. 886–893.
[170] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, B. Girod, CHoG: Compressed histogram of gradients a low bit-rate feature descriptor, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, vol. 20(25), 2009, pp. 2504–2511.
[171] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comp. Vis. 42 (3) (2001) 145–175.
[172] C.H. Lin, R.T. Chen, Y.K. Chan, A smart content-based image retrieval system based on color and texture feature, Image Vis. Comput. 27 (6) (2009) 658–665.
[173] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2564–2571.
[174] E. Rosten, R. Porter, T. Drummond, Faster and better: a machine learning approach to corner detection, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 105–119.
[175] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: BINARY robust independent elementary features, in: Computer Vision-ECCV, Springer, Berlin Heidelberg, 2010, pp. 778–792.
[176] C.H. ILampert, M.B. Blaschko, T. Hofmann, Beyond sliding windows: Object localization by efficient subwindow search, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008, pp. 1–8.
[177] X. Yang, K.T. Cheng, Accelerating surf detector on mobile devices, in: Proceedings of the 20th ACM International Conference on Multimedia, 2012, pp. 569–578.
[178] A. Torralba, R. Fergus, Y. Weiss, Small codes and large databases for recognition, in: CVPR, 2008, pp. 1–8.
[179] M. Muja, D.G. Lowe, Fast matching of binary features, in: IEEE Ninth Conference on Computer and Robot Vision (CRV), 2012, pp. 404–410.
[180] L. Zhuo, B. Cheng, J. Zhang, A comparative study of dimensionality reduction methods for large-scale image retrieval, Neurocomputing 141 (2014) 202–210.
[181] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1704–1716.
[182] L.J.P. Maaten, E.O. Postma, H.J. Herik, Dimensionality reduction: a comparative review, J. Mach. Learn. Res. 10 (1–41) (2009) 66–71.
[183] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[184] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[185] T.S. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, Advances in Neural Information Processing Systems, vol. 11, MIT Press, 1999.
[186] F. Perronnin, C.R. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[187] H. Jegou, M. Douze, C. Schmid, P. Perez, Aggregating local Descriptors into a compact image representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.
[188] P.N. Belhumeur, J.P. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[189] Y. Rahulamathavan, R.C.W. Phan, J.A. Chambers, D.J. Parish, Facial expression recognition in the encrypted domain based on local fisher discriminant analysis, IEEE Trans. Affect. Comput. 4 (suppl. 1) (2013) 83–92.
[190] M. Gashler, D. Ventura, T. Martinez, Iterative non-linear dimensionality reduction with manifold sculpting, in: NIPS, vol. 8, 2007, pp. 513–520.
[191] J.B. Tenenbaum, V. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[192] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[193] X. Niyogi, Locality preserving projections, Neural Information Processing Systems, vol. 16, MIT, 2004, p. 153.
[194] C.S. Anan, R. Hartley, Optimised KD-trees for fast image descriptor matching, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[195] V.P.R. Subramanyam, S.K. Sett, Image retrieval system using R-tree self-organizing map, Data Knowl. Eng. 61 (3) (2007) 524–539.
[196] T. Skopal, J. Lokoc, New dynamic construction techniques for M-tree, J. Discr. Algor. 7 (1) (2009) 62–77.
[197] X. Zhang, Z. Li, L. Zhang, W. Ma, H.Y. Shum, Efficient indexing for large scale visual search, in: IEEE 12th Conference on Computer Vision, 2009, pp. 1103–1110.
[198] M. Moro, D. Zhang, V.J. Tsotras, Hash-based Indexing, in: Encyclopedia of Database Systems, Springer, US, 2009, pp. 1289–1290.
[199] P. Indyk, R. Motwani, Approximate nearest neighbor: towards removing the curse of dimensionality, in: 30th Annual ACM Symposium on Theory of Computing, 1998, pp. 604–613.
[200] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Advances in Neural Information Processing Systems, 2009, pp. 1753–1760.
[201] L. Pauleve, H. Jegou, L. Amsaleg, Locality sensitive hashing: a comparison of hash function types and querying mechanisms, Pattern Recogn. Lett. 31 (11) (2010) 1348–1358.
[202] R. Salakhutdinov, G. Hinton, Semantic hashing, Int. J. Approx. Reason. 50 (7) (2009) 969–978.
[203] J. Shao, F. Wu, C. Ouyang, X. Zhang, Sparse spectral hashing, Pattern Recogn. Lett. 33 (3) (2012) 271–277.
[204] J.P. Heo, Y. Lee, J. He, S.F. Chang, S.E. Yoon, Spherical hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2957–2964.
[205] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[206] X. Chen, C. Zhang, SC. Chen, M. Chen, A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval, in: Seventh IEEE International Symposium on Multimedia, vol. 8, 2005, pp. 12–14.
[207] W. Liu, W. Xu, L. Li, W. Wang, Applying visual attention computational model and latent semantic indexing to image retrieval, in: 4th IEEE Conference on Industrial Electronics and Applications, ICIEA, vol. 2667(2671), 2009, pp. 25–27.
[208] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw. 10 (3) (1999) 626–634.
[209] M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[210] G. Bordogna, G. Pasi, Soft clustering for information retrieval applications, WIREs Data Min. Knowl. Discov. 1 (2011) 138–146.
[211] M. Bilenko, S. Basu, R.J. Mooney, Integrating constraints and metric learning in semi-supervised clustering, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, 2004, p. 11.
[212] C. Papagiannopoulou, V. Mezaris, Concept-based image clustering and summarization of event-related image collections, in: Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia, ACM, 2014, pp. 23–28.
[213] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[214] J. Kittler, F. Roli, Multiple classifier systems for robust classifier design in adversarial environments, Int. J. Mach. Learn. Cybernet. 1 (1–4) (2010) 27–41.
[215] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[216] Y. Freund, RE. Schapire, Experiments with a new boosting algorithm, in: ICML, vol. 96, 1996, pp. 148–156.
[217] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[218] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1619–1630.
[219] S. Kotsiantis, Combining bagging, boosting, rotation forest and random subspace methods, Artif. Intell. Rev. 35 (3) (2011) 223–240.
[220] I.A. Basheer, M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, J. Microbiol. Meth. 43 (1) (2000) 3–31.
[221] S.B. Park, J.W. Lee, S.K. Kim, Content-based image classification using a neural network, Pattern Recog. Lett. 25 (3) (2004) 287–300.
[222] M. Ghiassi, C. Burnley, Measuring effectiveness of a dynamic artificial neural network algorithm for classification problems, Expert Syst. Appl. 37 (4) (2010) 3118–3128.
[223] H. Yoon, C.S. Park, J.S. Kim, J.G. Baek, Algorithm learning based neural network integrating feature selection and classification, Expert Syst. Appl. 40 (1) (2013) 231–241.
[224] G.D. Wu, P.H. Huang, A vectorization-optimization-method-based type-2 fuzzy neural network for noisy data classification, IEEE Trans. Fuzzy Syst. 21 (1) (2013) 1–15.
[225] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technomterics 12 (1970) 55–67.
[226] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. Ser. B (Methodol.) (1996) 267–288.
[227] Li. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3 (2014) e2.
[228] A. Krizhevsky, I. Sutskever, GE. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[229] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1891–1898.
[230] H. Azizpour, A.S. Razavian, J. Sullivan, A. Maki, S. Carlsson, From Generic to Specific Deep Representations for Visual Recognition, 2014, Available from: arXiv:1406.5774.
[231] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Computer Vision and Pattern Recognition, 2014, Available from arXiv:1412.2306.
[232] <http://googleresearch.blogspot.co.uk/2014/11/a-picture-is-worth-thousand-coherent.html> (access May 2015).
[233] W. Bian, D. Tao, Biased discriminant euclidean embedding for content-based image retrieval, IEEE Trans. Image Process. 19 (2) (2010) 545–554.
[234] L. Zhang, L. Wang, W. Lin, Semisupervised biased maximum margin analysis for interactive image retrieval, IEEE Trans. Image Process. 21 (4) (2012) 2294–2308.
[235] X. Zhou, T. Huang, Small sample learning during multimedia retrieval using biasmap, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 11–17.
[236] L. Zhang, L. Wang, W. Lin, Generalized biased discriminant analysis for content-based image retrieval, IEEE Trans. Syst., Man, Cybernet., Part B: Cybernet. 42 (1) (2012) 282–290.
[237] E. Rashedi, H. Nezamabadi-Pour, S. Saryazdi, Long term learning in image retrieval systems using case based reasoning, Eng. Appl. Artif. Intell. 35 (2014) 26–37.
[238] C.C. Lai, Y.C. Chen, A user-oriented image retrieval system based on interactive genetic algorithm, IEEE Trans. Instrument. Measur. 60 (10) (2011) 3318–3325.
[239] C.D. Ferreira, J.A. Santos, R.S. Torres, M.A. Gonçalves, R.C. Rezende, W. Fan, Relevance feedback based on genetic programming for image retrieval, Pattern Recog. Lett. 32 (1) (2011) 27–37.
[240] S.R. Bulo, M. Rabbi, M. Pelillo, Content-based image retrieval with relevance feedback using random walks, Pattern Recog. 44 (9) (2011) 2109–2122.
[241] J.H. Su, W.J. Huang, P.S. Yu, V.S. Tseng, Efficient relevance feedback for content-based image retrieval by mining user navigation patterns, IEEE Trans. Knowl. Data Eng. 23 (3) (2011) 360–372.
[242] D. Keim, G. Andrienko, J.D. Fekete, C. Görg, J. Kohlhammer, G. Melançon, Visual Analytics: Definition, Process, and Challenges, Springer, Berlin Heidelberg, 2008, pp. 154–175.
[243] A. Kumar, F. Nette, K. Klein, M. Fulham, J. Kim, A visual analytics approach using the exploration of multi-dimensional feature spaces for content-based medical image retrieval, IEEE J. Biomed. Health Inf. 99 (2014) 2168–2194.
[244] J.J. Thomas, K.A. Cook, A visual analytics agenda, Comp. Graph. Appl., IEEE 26 (1) (2006) 10–13.
[245] A. Hiroike, Y. Musha, A. Sugimoto, Y. Mori, Visualization of information spaces to retrieve and browse image data, in: Visual Information and Information Systems, Springer, Berlin Heidelberg, pp. 155–163.
[246] J.F. Rodrigues, L.A.S. Romani, A.J.M. Traina, C. Traina, Combining visual analytics and content based data retrieval technology for efficient data analysis, in: 14th International Conference Information Visualisation, vol. 61(67), 2010, pp. 26–29.
[247] M. Tory, T. Moller, Human factors in visualization research, IEEE Trans. Visual. Comp. Graph. 10 (1) (2004) 72–84.
[248] T.M. Deserno, S. Antani, R. Long, Ontology of gaps in content-based image retrieval, J. Dig. Imaging 22 (2) (2009) 202–215.
[249] M. Wilson, Search-User Interface Design, Morgan & Claypool Publishers, 2011.

[250] D. Kelly, Methods for evaluating interactive information retrieval systems with users, Found. Trends Inf. Ret. 3 (1–2) (2009) 1–224.
[251] P. Ingwersen, K. Järvelin, The Turn: Integration of Information Seeking and Retrieval in Context, vol. 18, Springer Science & Business Media, 2006.
[252] A. Kumar, J. Kim, L. Bi, M. Fulham, D. Feng, Designing user interfaces to enhance human interpretation of medical content-based image retrieval: application to PET-CT images, Int. J. Comp. Assis. Radiol. Surg. 8 (6) (2013) 1003–1014.
[253] A. Ralescu, Generalization of the hamming distance using fuzzy sets, JSPS Senior Res. Fellowship, Lab. Math. Neurosci., Brain Sci. Inst., RIKEN, 2003.
[254] Y. Rubner, C. Tomasi, L.J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comp. Vis. 40 (2) (2000) 99–121.
[255] J. Puzicha, T. Hofmann, J.M. Buhmann, Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings, 1997, pp. 267–272.
[256] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, Pattern Recog. Lett. 19 (3–4) (1998) 255–259.
[257] M.L. Fernandez, G. Valiente, A graph distance metric combining maximum common subgraph and minimum common supergraph, Pattern Recog. Lett. 22 (6–7) (2001) 753–758.
[258] J. Raymond, E. Gardiner, P. Willett, RASCAL: calculation of graph similarity using maximum common edge subgraphs, Comp. J. 45 (6) (2002) 631–644.
[259] H. Bunke, On a relation between graph edit distance and maximum common subgraph, Pattern Recog. Lett. 18 (8) (1997) 689–694.
[260] Z. Zeng, A.K.H. Tung, J. Wang, J. Feng, L. Zhou, Comparing stars: on approximating graph edit distance, in: Proceedings of PVLDB, 2009, pp. 25–36.
[261] L. Yang, R. Jin, Distance Metric Learning: A Comprehensive Survey, vol. 2, Michigan State University, 2006, pp. 1–51.
[262] Z. Wang, Y. Hu, L.T. Chia, Learning image-to-class distance metric for image classification, ACM Trans. Intell. Syst. Technol. (TIST) 4 (2) (2013) 1–22.
[263] J. Wu, JM. Rehg, Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: IEEE 12th International Conference on Computer Vision, 2009, pp. 630–637.
[264] L. Wu, S.C. Hoi, Enhancing bag-of-words models with semantics-preserving metric learning, IEEE Multim. 18 (1) (2011) 24–37.
[265] SC. Hoi, W. Liu, SF. Chang, Semi-supervised distance metric learning for collaborative image retrieval and clustering, ACM Trans. Multim. Comput., Commun., Appl. (TOMCCAP) 6, 3(18) (2010) 1–25.
[266] Y. Zhang, D.Y. Yeung, Transfer metric learning by learning task relationships, in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 1199–1208.
[267] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, 2002.
[268] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, IEEE Com. Vis. Pattern Recog. (CVPR) (2009) 248–255.
[269] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features From Tiny Images, Computer Science Department, University of Toronto, Tech. Rep, 2009.
[270] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Comp. Vis. Image Understand. 106 (1) (2007) 59–70.
[271] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, 2007.
[272] <https://www.flickr.com/> (access May 2015).
[273] Z.W. James, J. Li, G. Wiederhold, SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (9) (2001) 947–963.
[274] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, Int. J. Comp. Vis. 77 (1–3) (2008) 157–173.
[275] A. Torralba, R. Fergus, W.T. Freeman, 80 million tiny images: a large data set for nonparametric object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1958–1970.
[276] J. Xiao, J. Hays, KA. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3485–3492.
[277] Y. LeCun, F.J. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: CVPR, vol. 2, 2004, pp. 97–104.
[278] TS. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, p. 48.
[279] G. Patterson, C. Xu, H. Su, J. Hays, The SUN attribute database: beyond categories for deeper scene understanding, Int. J. Comp. Vis. 108 (1–2) (2014) 59–81.
[280] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes (VOC) challenge, Int. J. Comp. Vis. 88 (2) (2010) 303–338.
[281] J. Kalpathy-Cramer, A.G.S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, H. Müller, Evaluating performance of biomedical image retrieval systems-an overview of the medical image retrieval task at ImageCLEF 2004–2013, Computer. Med. Imag. Graph. 39 (2015) 55–61.
[282] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, 2014, Available from arXiv:1409.0575.
[283] S.A. Chatzichristofis, C. Iakovidou, Y.S. Boutalis, E. Angelopoulou, Mean Normalized Retrieval Order (MNRO): a new content-based image retrieval performance measure, Multim. Tools Appl. (2012) 1–32.
[284] Euripides G.M. Petrakis, Design and evaluation of spatial similarity approaches for image retrieval, Image Vision Comput. 20 (1) (2002) 59–76.
[285] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1704–1716.
[286] B. Fan, F. Wu, Z. Hu, Aggregating gradient distributions into intensity orders: a novel local image descriptor, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2377–2384.
[287] F. Perronnin, Y. Liu, J. Sánchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3384–3391.
[288] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A Deep Convolutional Activation Feature for Generic Visual Recognition, 2013, Available from: arXiv:1310.1531.
[289] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV, Springer International Publishing, 2014, pp. 818–833.
[290] G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking, J. Mach. Learn. Res. 11 (2010) 1109–1135.
[291] S. Zhang, J. Huang, H. Li, D.N. Metaxas, Automatic image annotation and retrieval using group sparsity, IEEE Trans. Syst., Man, Cybernet., Part B: Cybernet. 42 (3) (2012) 838–849.
[292] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, D. Metaxas, Automatic image annotation using group sparsity, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3312–3319.
[293] S. Gao, L.-T. Chia, I.W.-H. Tsang, Multi-layer group sparse coding for concurrent image classification and annotation, in: IEEE CVPR, 2011, pp. 2809–2816.
[294] Y. Yang, Y. Yang, Z. Huang, H.T. Shen, F. Nie, Tag localization with spatial correlations and joint group sparsity, in: IEEE CVPR, 2011, pp. 881–888.
[295] G. Guo, A. Lai, A survey on still image based human action recognition, Pattern Recog. 47 (10) (2014) 3343–3361.
[296] F. Christiane, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[297] MJ. Huiskes, MS. Lew, The MIR Flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.
[298] F.F. Faria, A. Veloso, H.M. Almeida, E. Valle, R.S. Torres, M.A. Gonçalves, W.M. Jr., Learning to rank for content-based image retrieval, in: Proceedings of the International Conference on Multimedia Information Retrieval, ACM, 2010, pp. 285–294.
[299] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, J. Luo, Efficient manifold ranking for image retrieval, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 525–534.
[300] B. Siddiquie, R.C. Feris, L.S. Davis, Image ranking and retrieval based on multi-attribute queries, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 801–808.
[301] D.C.G. Pedronette, J. Almeida, R.S. Torres, A scalable re-ranking method for content-based image retrieval, Inf. Sci. 265 (2014) 91–104.
[302] T. Mei, Y. Rui, S. Li, Q. Tian, Multimedia search reranking: a literature survey, ACM Comput. Surv. (CSUR) 2 (3) (2014) 1–36.
[303] I. Ruthven, Interactive information retrieval, Ann. Rev. Inf. Sci. Technol. 42 (2008) 43–91.
[304] D. Kelly, Methods for evaluating interactive information retrieval systems with users, J. Found. Trends Inf. Ret. 3 (1–2) (2009) 1–224.
[305] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medical applications—clinical benefits and future directions, Int. J. Med. Inf. 73 (1) (2004) 1–23.
[306] A. Kumar, J. Kim, W. Cai, M. Fulham, D. Feng, Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data, J. Dig. Imag. 26 (6) (2013) 1025–1039.
[307] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1–3) (2008) 157–173.
[308] J. Li, J. Wang, Real-time computerized annotation of pictures, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 985–1002.