Neurocomputing 452 (2021) 675–689

Contents lists available at ScienceDirect

Neurocomputing

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

Recent developments of content-based image retrieval (CBIR)

Xiaoqing Li, Jiansheng Yang, Jinwen Ma *

Department of Information and Computational Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 10 April 2020
Revised 4 July 2020
Accepted 21 July 2020
Available online 4 January 2021

Keywords:
Content-based image retrieval
Image representation
Database search
Computer vision
Big data
Deep learning

ABSTRACT

With the development of Internet technology and the popularity of digital devices, Content-Based Image Retrieval (CBIR) has been quickly developed and applied in various fields related to computer vision and artificial intelligence. Currently, it is possible to retrieve related images effectively and efficiently from a large scale database with an input image. In the past ten years, great efforts have been made for new theories and models of CBIR and many effective CBIR algorithms have been established. In this paper, we present a survey on the fast developments and applications of CBIR theories and algorithms during the period from 2009 to 2019. We mainly review the technological developments from the viewpoint of image representation and database search. We further summarize the practical applications of CBIR in the fields of fashion image retrieval, person re-identification, e-commerce product retrieval, remote sensing image retrieval and trademark image retrieval. Finally, we discuss the future research directions of CBIR with the challenge of big data and the utilization of deep learning techniques.

1. Introduction

With the development of Internet technology and the popularity of digital devices, it is easy and convenient to take a photo or get an image on any object we are interested in. In fact, there are a huge number of images generated in our daily life. So, these images can be utilized to improve the performance of information processing and make our life more intelligent and convenient. Actually, Content-Based Image Retrieval (CBIR) technique can retrieve related images from a database with an input image of the object or content we are interested in, which is widely used in various fields of computer vision and artificial intelligence. Face retrieval [1] can help police and other security personnels catch suspects more quickly. In online shopping, commodity image retrieval [2] can help customers find their favorite commodities. Building retrieval [3] from the map can help us locate more accurately and reduce the possibility of getting lost. Clothe retrieval [4] can help consumers buy the clothes they want. Medical image retrieval [5] can help doctors make medical diagnosis more effectively and so on. As a matter of fact, many effective CBIR systems have been developed and applied for those practical applications in recent years.

For a CBIR system, there are two major mechanisms or components which are respectively image representation for image indexing and similarity measure for database search. Feature vector or image representation is expected to be discriminative so as to distinguish images. More importantly, it is also expected to be invariant to certain transformations. Based on image representation, the similarity measure between two images should reflect the relevance in semantics. These two connected components are crucial to the retrieval performance and the existing algorithms of CBIR can be categorized according to their contributions to these two components. In fact, it is still challenging to get an accurate retrieval image from a large-scale database. The greatest challenge is the semantic gap between the high-level meaning of the image and its low-level visual features [6]. To narrow this semantic gap, extensive efforts have been made from both academia and industry. Consequently, CBIR has been witnessed to make great advances in recent years. For example, Google and Baidu are popular search engines which can search the related image by any image. Some e-commerce sites like Alibaba, Amazon and eBay have similar commodity search functions. Social platforms like Pinterest have similar content recommendation functions.

Recently, there are already some surveys related to CBIR. Zheng et al. [7] surveyed the image search from 2006 to 2016 based on Scale-Invariant Feature Transform (SIFT) and Convolutional Neural Network (CNN). Radenovic et al. [8] further surveyed the related search methods from the perspective of Oxford and Paris datasets. Zhou et al. [9] surveyed the CBIR researches in the past decade after 2003. However, there are some deficiencies in these surveys. On one hand, they did not include the latest researches from 2017 to 2019 during which the image retrieval technology has developed rapidly with the challenge of big data and the utilization of

* Corresponding author.
E-mail address: jwma@math.pku.edu.cn (J. Ma).

https://doi.org/10.1016/j.neucom.2020.07.139
0925-2312/

deep learning techniques. In fact, there are many new image retrieval algorithms which are worth being summarized. On the other hand, they only focused on the technological developments, without the in-depth summary of practical applications. In this survey, we focus on both the technological developments and practical applications of CBIR from 2009 to 2019.

The main contributions of this paper lie in three aspects:

- This paper is the first to review and classify the technological developments from the viewpoint of image representation and database search during the period from 2009 to 2019.
- We further summarize the practical applications of CBIR in the fields of fashion image retrieval, person re-identification, e-commerce product retrieval, remote sensing image retrieval and trademark image retrieval during the period from 2009 to 2019.
- We analyze and discuss the future research directions of CBIR with the challenge of big data and the utilization of deep learning techniques

The rest of this paper is organized as follows. Section 2 presents a general pipeline of CBIR. Section 3 and Section 4 then review the CBIR developments of image representation and database search in recent years, respectively. The practical applications of CBIR are further summarized in Section 5. Furthermore, we outline the future potential directions of CBIR in Section 6. We finally make a brief conclusion in Section 7.

2. General flowchart overview

We begin to introduce the general framework of CBIR system which can be further divided into an off-line subsystem and an online subsystem, as shown in Fig. 1. In the off-line subsystem, every image is coded by its extracted feature vector as the index in the retrieval database. In the on-line subsystem, after a query image is inputted, its feature vector is extracted in the same way as those of the images in the retrieval dataset. Then, we use this feature vector to score all the possible images in the database with a similarity measure. Those images with a higher score than a predefined threshold are selected to be further refined by enhancing the visual context in contrast to the original query. Finally, these images in the descent order of the rerank score are considered as the probability-ordered results or outputs of the retrieval system. In this framework, the feature-based image representation is fundamental for the dataset indexing with the help of certain similarity measure. From a technological viewpoint, the CBIR system is based on image representation and database search. So, we can survey the CBIR researches from the developments of image representation and database search, respectively.

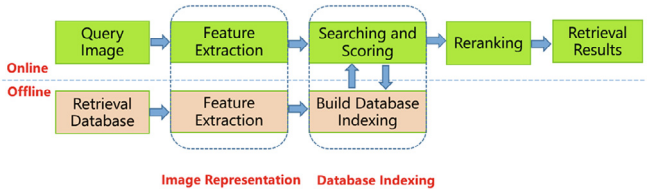

Fig. 1. The general framework of CBIR. According to two different ways of information processing, the CBIR system is divided into online and offline subsystems, but they share the same feature extraction block.

3. Image representation

For CBIR, the key step is the image representation that extracts the critical features from a given image and then transforms them into a fix-sized vector (so called feature vector). In general, the extracted features can be divided into three main categories: conventional features, classification CNN features, and retrieval CNN features. In this section, we summarize the methods of image representation for CBIR according to these three feature categories in the following subsections, respectively. For clarity, the hierarchical categories of image representation based methods are shown as Fig. 2.

3.1. Conventional feature based methods

Conventional features here refer to the features which are not extracted through any CNN methods. They are mainly used in the early CBIR systems, but also have developed remarkably in certain ways in recent years. These features are heuristically designed and can be further categorized into the global and local features. The global features are usually extracted from the color [10], shape [11], texture [12], and structure of an image,respectively, and then transformed into a holistic representation. Certainly, these multi-type global features can be further combined together for image retrieval. Li et al. [13] actually developed a two-phase generative/discriminative learning algorithm to combine color, texture, and structure features together for image retrieval. Specifically, the generative phase normalized the lengths of various descriptions of the images, while the discriminative phase recognized which images contain the target object. In fact, this algorithm can combine any number of different feature types without any modeling

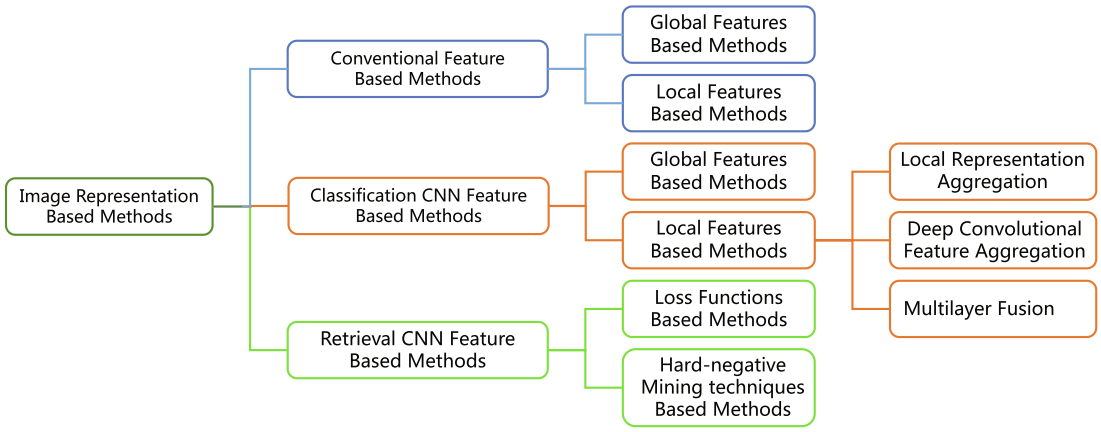

Fig. 2. The hierarchical categories of image representation based methods.

assumptions. Berman et al. [14] presented a set of operations for combining features and proposed a Flexible Image Database System (FIDS) in which the gridded distance measures and combining operations can also combine any group of multiple types of features. Moreover, Zhao et al. [15] adopted the Sparse Representation (SR) into image retrieval. They firstly explored the difference and complementarity between the forward and backward SRs, and then built a novel semi-supervised learning model called cooperative sparse representation, which effectively combines them to improve image annotation performance. As a result, this SR based method achieved a good result on image retrieval. Although these methods are suitable for duplicate detection in a large-scale image database because of their compact expression, they may not work well when the target images involve some background clutters.

On the other hand, one of the most famous local features is SIFT, which mainly involves two steps: interest point detection and local region description. In recent years, many local feature extraction methods are the extensions of SIFT. For example, Zhou et al. [16] developed the binary signature of the SIFT descriptor with two median thresholds determined by the original descriptor itself. Moreover, a new indexing scheme BSIFT for CBIR is established with this binary SIFT [17]. Furthermore, on the basis of SIFT, the edge is also added into the feature descriptor to establish Edge-SIFT [18] and so on. Apart from the feature extraction methods of image key points like SIFT, some local features methods extract the features on the dense grids, possibly at multiple scales independently of the image content [19]. In fact, a variety of local descriptors have been developed in recent years. Since these method have different kinds of superiority as claimed, it is rather difficult to select the best one for a retrieval task. Nevertheless, Madeo et al. [20] made a comparative analysis among some typical local descriptors from three aspects: speed, compactness, and discrimination.

When the sample image changes greatly with a large set of some kind local features through the dataset, it is often necessary to aggregate these local features into a vector representation with a fixed length for the subsequent database search via the similarity comparison of a query against all the database images. Most of these aggregation schemes need to make a clustering analysis on these local features to obtain a codebook of the centers of the obtained clusters. According to this codebook, the original feature vector can be aggregated in different ways. Sivic et al. [21] proposed the Bag-of-Words (BoW) method which uses the k-means algorithm to create the codebook. Then the clustering center nearest to the feature point is used to replace the feature point. Usually, this aggregation scheme can lose certain detailed information and the generated BoW vector is very sparse. Perronnin et al. [22] further proposed the Fisher Vector (FV) which aggregates local descriptors using the Gaussian Mixture Model(GMM). Actually, GMM can be used for clustering analysis, and it considers the distance from the feature point to each cluster center directly. In the FV method, each feature point is represented by a linear combination of all cluster centers. And, this aggregation scheme also loses some information in the process of GMM modeling. Based on BOW and FV schemes, Jegou et al. [23] proposed Vector of Locally Aggregated Descriptors (VLAD) scheme. On one hand, like BOW, VLAD only considers the cluster center closest to the feature point, and saves the distance from each feature point to the cluster center closest to it. On the other hand, like FV, VLAD considers the value of each dimension of the feature point that has a more detailed description of the local information of the image. More importantly, the VLAD feature has no loss of information. Some other works [24–28] are the improvement or extension of VLAD, which have been demonstrated by many experiments. Robust Visual Descriptor (RVD) proposed by Husain [29] combined the rank-based multi-assignment with robust accumulation to reduce the impact of outliers which is a relatively new investigation.

3.2. Classification CNN feature based methods

Because of the limited representation ability of conventional features and the breakthrough of image processing via deep neural networks, CNN based image retrieval has developed quickly in recent years. Since CNN has made a big breakthrough on image classification, many researchers attempt to use the CNN features trained with a classification task for CBIR. In this subsection, we make a summary of those methods with the classification CNN features.

In the same way as above, CNN features can be also categorized into two types: the global and local ones. The features from the deep fully connected layers of CNN are essentially deep global features which describe the overall semantic information of the image. There exist certain methods based on deep global features. Babenko et al. [30] took the activation of fully connected layers where CNN is finetuned on the dataset that is relevant to the test set as the global descriptors followed by dimensionality reduction. However, the cost of making labeled training data is expensive, so some works use the off-the-shelf networks only pre-trained on ImageNet. Babenko et al. [30] showed that using the 7th fully connected layer (fc7) features can have a better retrieval effect than using the 8th fully connected layer (fc8) features. This is because that the higher layer features are intended to perform the classification task on the pre-training dataset, while the lower layer features have better generalization capabilities for the other datasets. These methods can produce compact embedding features to enable fast similarity computation in the filtering step, but may lack a description of the details of the image, which is not so significant for image retrieval. Therefore, there are more and more works focusing on local features of CNN.

According to the features in image representation extraction process, local CNN feature-based methods can be further divided into three categories: local representation aggregation, deep convolutional feature aggregation, and multi-layer fusion. Usually, these methods use the off-the-shelf networks pre-trained only on ImageNet as feature extractors.

We begin to consider the methods of local representation aggregation. Actually, they first try to extract a series of local regions from an input image, and then feed these local regions to the network and generate the corresponding partial image representations. So, these partial image representations are aggregated into a compact image representation by a specific aggregation method. According to the difference of extracting local regions, we further classify this kind of methods into three categories: local area extraction based on the sliding window, region of interest detection, and local area extraction based on region proposal. The first type of methods usually slide on the input image with a series of different sliding windows to create a partial area. Razavian et al. [31] used 4 different sizes of sliding windows and fc7 features as a partial representation of the image. Gong et al. [32] proposed the Multi-scale Orderless Pooling (MOP) algorithm to embed and pool the CNN fully connected activation of image patches of an image at different scale levels, then used the VLAD to aggregate these partial image representations. The second type of methods implement some specific detection algorithms to extract the regions of interest in the input images. For example, Patch-Convolutional Kernel Network (Patch-CKN) [33] utilized the Hessian-affine detector to extract regions of interest in input images. Mopuri et al. [34] leveraged a detector trained for robust landmark localization to produce more efficient regional search systems. The third type of methods are based on region proposal and utilize some specific unsuper-

vised candidate region generation algorithms to obtain local candidate regions that may contain targets. Mopuri et al. [35] used the selective search to extract 2000 local regions and fed these regions to the network, and then computed an image-level representation by the max-pooling of the generated fc7 features. Complementary CNN and SIFT (CCS) [36] used the EdgeBox to extract 100 local regions per image and fed these regions to the network, and then used the VLAD to aggregate these CNN features. All of the above methods require multiple feedforward networks, so their efficiencies become the bottleneck.

Deep convolutional feature aggregation methods are based on the fact that the deep convolution feature can be regarded as a description of the local area of an input image. These methods only feed the network once to generate the deep convolution features and aggregate them to get the final representation of the image, so the calculation efficiency is relatively high. The key problem of these methods is how to aggregate the deep convolution features. By considering whether there is the weighting mechanism in the aggregation, we can classify such methods into direct aggregation and weighted aggregation. The direct aggregation methods use some specific aggregation methods to aggregate the deep convolution features to obtain the final image representation. The aggregation methods can use the classic methods like BOW, VLAD, FV, max-pooling, and sum-pooling. Regional-Maximum Activation of Convolutions (R-MAC) [37] uniformly sampled and aggregated by max-pooling local regions in a convolutional feature map for considering region-wise information. Sum-Pooled Convolutional (SPoC) [38] showed that the sum-pooling method outperformed the max-pooling method when the final image representation was whitened. The retrieval performance was further improved when the Robust Visual Descriptor with Whitening (RVD-W) method [29] was used for the aggregation of CNN features. Iscen et al. [39] leveraged the multi-scale grids in conjunction with CNN features to enable query expansion via diffusion. The weighted aggregation methods use the strategy of direct aggregation to encode deep convolution features, and weight the deep convolution features according to the importance of each location feature. Selective Convolutional Descriptor Aggregation (SCDA) [40] proposed an unsupervised method for localizing the representative object while removing the noisy background, resulting in the improvement of fine-grained image retrieval. Kalantidis et al. [41] proposed the Crow that extended SPoC by introducing cross-dimensional weighting in aggregation of CNN features. Jimenez et al. [42] employed the Class Activation Maps (CAMs) for calculating semantic-aware spatial weights of a convolutional feature map. It was found by Part-based Weighting Aggregation (PWA) [43] that the different channels of the deep convolution feature corresponded to the response of different parts of the target. Some retrieval methods [44,45] adopt the attention mechanisms. DEep Local Feature (DELF) [44] adopted a learning-based attention network and used the attention network for densely weighting all points of a feature map. Kim et al. [45] proposed a simple, yet effective regional attention network, which weighted an attentive score of a region considering the global context.

Finally, we consider the methods of multi-layer fusion. In fact, deep CNN features are hierarchical, that is, they change from the low-level visual features to the high-level semantic features as the layers of CNN goes deeper. As discussed above, deep convolutional feature aggregation methods aim to complement the information of different layer features in deep neural networks to synthesize the invariance and discriminative ability of different layer features. We can further fuse these multi-layer features together. In fact, MultiScale-Regional Maximum Activation of Convolutions (MS-RMAC) [46] extracted the R-MAC features [37] separately from relu1-2 layer, relu2-1 layer, relu3-3 layer, relu4-3 layer and relu5-3 layer in VGG-16, and then weighted the resulted features into the final image representations. Seddati et al. [47] proposed a modified RMAC signature that combined the multi-scale and two-layer feature extraction mechanisms through feature selection. Region-Entropy based Multi-layer Abstraction Pooling (REMAP) [48] learned and aggregated a hierarchy of deep features from multiple CNN layers, and was trained in the way of end-to-end with a triplet loss.

3.3. Retrieval CNN feature based methods

While so many researches focused on the features extracted from the deep networks pre-trained for a classification task, it was later shown that the deep networks could be trained directly for the task of instance retrieval in an end-to-end manner [49,50]. In order to enforce intra-class discrimination and more fine-grained instance-level image representations, current researches mainly use deep metric learning. Deep metric learning aims to learn an embedding space, where the embedded vectors of similar samples are encouraged to be closer, while dissimilar ones are pushed apart from each other. Its key is to leverage a loss function that optimizes the ranking instead of classification and has a hard-negative mining technique that improves the quality of learned embedding space [51,52]. This class of approaches represents the current state-of-the-art in image retrieval [53–55].

There are some effective loss functions in deep metric learning, such as contrastive loss, triplet loss, triplet-center loss, quadruplet loss, lifted structure loss, N-pairs loss, binomial deviance loss, histogram loss, angular loss, distance weighted margin-based loss, and hierarchical triplet loss. Their common principle is to subsample a small set of images, verify that they locally comply with the ranking objective, perform a small model update if they do not, and repeat these steps until convergence. Retrieval methods based on retrieval CNN features employ above loss functions. Radenovic et al. [56] used contrastive loss to learn image representation, while Gordo et al. [49] introduced the triplet loss. Similarly, the N-pairs loss [57], Triplet-center loss [58] and quadruplet loss [59] were all effective on improving retrieval accuracy. Kim et al. [54] proposed a new triplet loss that allows the distance ratios in the label space to be preserved in the learned metric space. Recently, He et al. [60] introduced the Average Precision (AP) loss and demonstrated its outstanding results in the context of patch verification, patch retrieval and image matching. Inspired by this, Revaud et al. [53] directly optimized the global mean Average Precision (mAP) by leveraging recent advances in listwise loss formulations. Compared with existing losses, it can consider thousands of images simultaneously at each iteration, and also establishes a new state-of-the-art image representation on many standard retrieval benchmarks. Hard-negative mining techniques can remarkably improve the quality of learned embedding space in deep metric learning. In fact, a lot of work has been done in this area for CBIR. The lifted structured loss [51] considered all positive and negative pairs in a mini batch at each time by incorporating hard-negative mining functionality within itself. Kim et al. [54] designed a triplet mining strategy adapted to metric learning with continuous labels. In general, the hard-negative mining techniques such as semi-hard mining, smart mining, and distance weighted sampling, can be used in retrieval methods based on retrieval CNN features. Especially, Wang et al. [55] established a General Pair Weighting (GPW) framework, which cast the sampling problem of deep metric learning into a unified view of pair weighting through gradient analysis, providing a powerful tool for understanding recent pair-based loss functions.

4. Database search

With the effective image representation or feature vector, we can further establish a database indexing method for the CBIR system and search with the image indexes via a similarity measure. For clarity, this whole process is referred to as database search. Because the retrieval time is key to the performance of the CBIR system, database search is very important, especially in a large-scale image database. In fact, an efficient database search method can significantly accelerate the retrieval process and reduces memory usage substantially. Database search methods for high-dimensional feature vector are usually divided into two types: feature-direct database search and feature-utilized database search. In practical applications, however, database search methods often combine these two kinds of database search methods together for the performance optimization. For clarity, the hierarchical categories of database search methods are shown as Fig. 3.

4.1. Feature-direct database search

Feature-direct database search methods utilize the feature vector of each image as its index for database search, that is, the feature vector is directly the index for the original image. In fact, there exist certain efficient feature-direct database search methods for CBIR. Some methods speed up the retrieval process by the inverted file index techniques, while the other methods narrow the search roads by the hierarchical clustering or K-Dimensional Tree (KD Tree). The most commonly used search method of this kind is the Inverted File Index (IFI) [61]. So, we give a summary on the development of IFI for CBIR in recent years.

Inspired by the field of information retrieval, the IFI stores the mapping of unique word IDs to the document IDs in which the words occur. In the CBIR system, IFI is a compact representation of a sparse matrix whose rows and columns represent images and visual words, respectively. So, in the query phase, the database image containing the common visual word with the query image will participate in the calculation of similarity such as euclidean distance, cosine distance and so on, which greatly improves the time efficiency. IFI is the central component of many search systems [62,63] as it facilitates the faster and more scalable querying. Some methods have made certain improvements to the original IFI. Inverted Multi-Index (IMI) [64] uses the idea of product quantization to construct a multi-index structure to optimize the reordering search space. The index of the traditional IFI structure is in one-dimensional data, and the index of the IMI structure uses a multi-dimensional table. When using the IMI for retrieval, the returned candidate inverted list is shorter, and the candidate elements are closer to the query word, and the recall rate is higher. Zheng et al. [65] used the multiple Inverse Document Frequency (IDF) methods to adapt the correlation between multiple features and store the corresponding binary code of the feature into the inverted index, while Liu et al. [66] proposed the cross-indexes of original SIFT feature space and the binary SIFT space.

In order to improve retrieval accuracy, some methods embed semantic information in the inverted table. Karayev et al. [67] removed the irrelevant images in the inverted table based on SIFT features by semantic attributes, and inserted semantically related images, which greatly enhanced the distinguishing power of features in the index. Zhang et al. [68] proposed a method to co-index semantic attributes into inverted index generated by local features, which makes the index convey more semantic cues. There also are some improved algorithms to speed up the retrieval. Zheng et al. [69] proposed a Q index, which removed the unimportant features of the query image and only retrieved more important features of the inverted table based on the predefined feature scores. For parallel retrieval, Ji et al. [70] built distributed indexes on multiple servers and defined the index distribution problem as a learning problem to reduce search latency between servers.

In order to increase the recall rate, multiple quantizers are usually used on an image to get multiple indexes [71,72]. Xia et al. [71] used a collaborative index structure to optimize multiple quantizers simultaneously. Non-Orthogonal Inverted Multi-Index (NO-IMI) proposed by et al. [72], was a fast indexing method for massive deep feature data. The NO-IMI included two code tables, S and T, each of which contained K code words. S is a first-order code table and generated by clustering of raw data, while T is a second-order code table and generated by clustering the residual data between the original data and the centroid of each corresponding first-order cluster (the code word in S), which thus gets rid of any decomposition of orthogonal subspaces. Therefore, the NO-IMI provided more reasonable index cells with the centroids representing actual data distribution more accurately for massive deep CNN feature data indexing.

4.2. Feature-utilized database search

Feature-utilized database search methods reduce the computational complexity of the distance by mapping the high-dimensional floating-point feature vectors into the low-dimensional vectors or binary vectors. One widely used feature-utilized database search method is the hashing-based indexing which compresses an image into a series of hashing codes such that the search can be converted into a comparison of the Hamming distances among the hashing codes. So, it can reduce both the computational complexity and the storage cost. Generally speaking, the hashing-based methods can be divided into the data-independent methods [73] and the data-dependent methods [74–77]. Actually, the data-dependent methods can be further divided into the unsupervised hashing methods [74,75] and the supervised hashing methods [76–78]. The hierarchical structure of the hash-based methods is shown in Fig. 4.

The data-independent methods are based on the hash functions which are generated independently without any information of

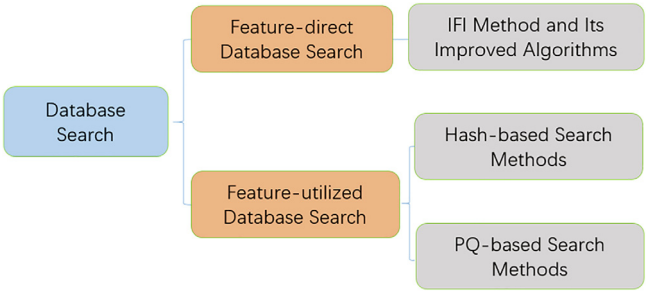

Fig. 3. The hierarchical categories of database search methods.

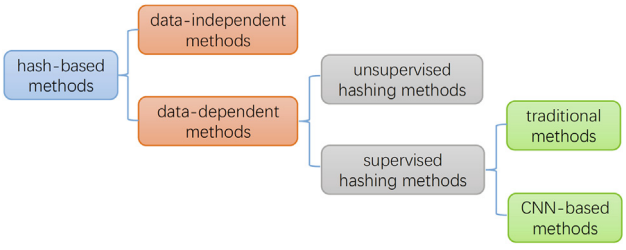

Fig. 4. The hierarchical categories of the hash-based methods where the methods in same color belong to the same category.

training data. The most representative one is the Locality-Sensitive Hashing (LSH) [73], which uses many random mapping hash functions to divide the feature space. When two feature vectors are similar, they have a higher probability of collision. Given a query image, a candidate list can be filtered based on the hash conflict, and then be reranked by the accurate distance calculation. There are lots of variants of this method to speed up the search process [79]. However, since they lack the information from the training dataset, their retrieval results are not satisfactory.

The unsupervised hashing methods only use the information of training dataset without label information to guide the training stage. Their representatives include the Isotropic Hashing (IsoH) [74], Scalable Graph Hashing (SGH) [80], Ordinal Embedding Hashing (OEH) [75] and so on. However, since there is no information from the training dataset to guide the training process, their results have certain limitations.

In the supervised hashing methods, they can be further divided into traditional supervised hashing methods and CNN-based supervised hashing methods. There exist some traditional supervised hashing methods, such as Fast supervised Hashing (FastH) [81], Supervised Discrete Hashing (SDH) [82], COlumn Sampling based DIscrete Supervised Hashing (COSDISH) [83] and Asymmetric Inner-product Binary Coding (AIBC) [84]. Compared with the unsupervised hashing, traditional supervised hashing methods can use the label of images to generate a higher precision score of hash codes. However, for streaming data, the hash models in the data-dependent methods should be modified to adapt the distribution of new coming data. With the continuous growth of data coming from the Internet, the online update of hashing on the massive social data becomes very time-consuming. To alleviate this issue, Ma et al. [85] proposed Hamming Subspace Learning (HSL), which was to generate a low-dimensional Hamming subspace from a high-dimensional Hamming space by selecting representative hash functions. HSL is effective to improve the speed of online updating and the performance of hashing in ceratin ways. Even so, these traditional supervised hashing methods have two obvious disadvantages: one is that their features extraction is independent of hash function learning such that the designed features might not be compatible with the hashing procedure; another one is that they use traditional features which can not include semantic information.

Recently, the CNN-based hashing methods become a hot spot of CBIR research. To start with, Convolutional Neural Network Hashing (CNNH) [86] pushed the deep hash algorithm based on CNN to the forefront. This is not an end-to-end training. Network In Network Hashing (NINH) [87] learned feature module and hash coding module simultaneously based on CNN. Although it was an end-to-end training, the accuracy of feature learning was not good enough. The Compact Root Bilinear CNN (CRB-CNN) [76] used the integrated network model to obtain better semantic features. Conjeti et al. [88] and Cheng et al. [78] both proposed the residual hash architecture to reduce the storage capacity of the computer and improved the retrieval efficiency. Deep Semantic Ranking Hashing (DSRH) [89] used a network structure similar to DeepID2 and directly let the network learn to rank. Similarly, Shi et al. [77] proposed a deep ranking hash for retrieval and classification tasks. Inspired by the online training strategy of Deep Supervised Hashing (DSH) [90] and taking the advantage of richer information on using triplet labels, Zhou et al. [91] utilized the triplet loss function to enhance the DSH algorithm for learning the compact binary codes. In addition, Li et al. [92] proposed a novel hash code generation method based on CNN, namely the Piecewise Supervised Deep Hashing (PSDH) method, which directly uses a latent layer data and the output layer result of the classification network to generate a two-segment hash code for every input image. In fact, PSDH performed excellently in the search of pictures with similar features. Most of these methods are of end-to-end training where their features extraction is dependent to hash function learning. They have significant retrieval accuracy and can save retrieval time. So, CNN-based hashing is getting more and more attention.

Another widely used feature-utilized database search method of CBIR is the Product Quantization (PQ) which decomposes the feature space into Cartesian products of multiple low-dimensional subspaces, and then quantizes each subspace separately. In the training phase, each subspace is clustered to obtain multiple centroids (quantizers), and the Cartesian product of all these centroids constitutes a dense partition of the whole space, which can ensure that the quantization error is relatively small. After the quantization learning, for a given query vector, the asymmetric distance of the query vector and each vector in the database can be calculated by looking up the table.

Generally, PQ is the best choice to generate a large codebook at very low memory storage and time cost. Because the hashing methods lack the accuracy of feature restoration, product quantization methods for minimizing the quantization error can achieve the superior accuracy over the hashing methods in some cases [93]. Like the supervised hashing methods, they can be also divided into traditional product quantization methods and CNN-based product quantization methods. For clarity, the hierarchical categories of the PQ-based methods are shown in Fig. 5. The representative traditional product quantization methods are the Product Quantization (PQ) [94], Optimized Product Quantization (OPQ) [95] and Composite Quantization (CQ) [96]. In fact, PQ [94] and OPQ [95] firstly split the whole feature space into many subspaces and then perform a similar algorithm on each subspace separately. CQ [96] learns enough codebooks using the same strategy as OPQ, but the dimension of its code words is equal to that of the original features. These traditional product quantization methods can generate large codebooks with low memory storage and time cost.

Recently, CNN-based product quantization methods have developed rapidly. They used end-to-end CNNs to perform the image feature learning and quantization together. Deep quantization network [97] is the first deep learning structure that learns and quantizes well separated image features. Then, deep visual-semantic quantization [93] uses CQ to quantize separated image features. Yu et al. [98] proposed a differentiable quantization method. However, the above methods have two disadvantages. First, they need to train many models if we want to get binary codes with different code lengths. Second, the decomposition of the high-dimensional vector space is tricky. To tackle these issues, both the Deep Recurrent Quantization (DRQ) [99] and Deep Progressive Quantization (DPQ) [100] made some efforts. The DRQ [99] used a deep quantization method to construct a codebook that could be used recurrently to generate sequential binary codes. Moreover, DPQ

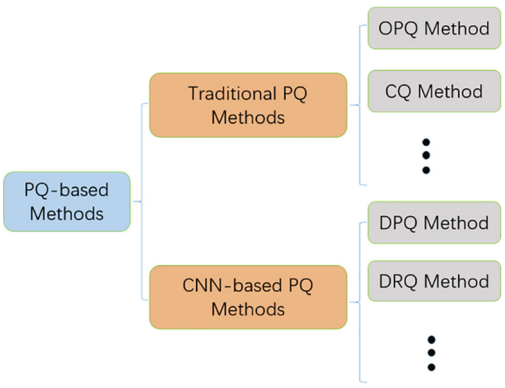

Fig. 5. The hierarchical categories of the PQ-based methods.

[100] proposed an alternative model to PQ for large scale image retrieval.

5. Practical applications

With the development of Internet technology and the popularity of digital devices, CBIR has been widely applied in many fields such as biomedicine, medicine, military, commerce, art and so on. In this section, we summarize some typical practical applications of CBIR in respect of objects to be retrieved, and categorize them into fashion image retrieval, person re-identification, e-commerce product retrieval, remote sensing image retrieval and trademark image retrieval, respectively.

5.1. Fashion image retrieval

The rapid development of clothing e-commerce and the increase of the amount of clothing image data on the Internet have made the Fashion Instance-level image Retrieval (FIR) be an increasing interest in computer vision. FIR is the task of finding fashion images similar to any query image, which satisfies the needs of users. Actually, FIR is mainly related to the cross-domain fashion image retrieval, which is to match two kinds of images, one kind is casually taken by users and another is professionally taken by the sellers. It plays an important role in the growing demands of online shopping, fashion recognition, and web-based recommendations. However, because the items of clothing are highly deformable, the viewpoints of images are severely changeful, and the shooting environment such as lighting and background are also various, the FIR has been considered as a challenging task.

In the last decade, many fashion datasets have been proposed to facilitate the FIR research. Table 1 gives a concise comparison among them at the view of the numbers of images, categories, pairs and the year they were released.

With the availability of large-scale fashion datasets, many FIR methods based on deep learning were proposed and worked well. These methods employed the advanced techniques of deep learning to enhance the retrieval performance. Some FIR methods adopt various attention mechanisms by using the advances of metric learning. Gajic et al. [106] paid much attention to improving training process and inference time. They stressed the importance of proper training of simple architecture, used the triplet loss to train the network, and adapted the general models to the specific task. Visual Attention Model (VAM) [107] trained a two-stream network with an attention branch and a global convolutional branch to form an end-to-end network structure, and then concatenates the generated vectors to optimize a standard triplet objective function. FashionNet [103] also trains network using a triplet loss. Zhao et al. [108] proposed an adversarial network for Hard Triplet Generation (HTG) to optimize the network ability in distinguishing similar examples of different categories as well as grouping varied examples of the same categories. Hard-aware Deeply Cascaded embedding (HDC) [109] combined a set of models with different complexities in cascaded mechanism to mine hard examples at multiple levels. Then, featured vectors from each sub-network were scaled by constant weights and concatenated to generate representations which are used for retrieval. Grid Search Network (GSN) [110] posited the training procedure as a search problem, focuses on locating matches for a reference query image in a grid containing both positive and negative images. Some FIR methods uses attribute modules [111,101,112,113]. Minchul et al. [113] aimed to achieve competitive performance on FIR. They proposed a novel method that converted a query into a representation with the desired attributes and introduced a new idea of attribute manipulation at the feature level. Some algorithms combine many techniques from deep learning. Park et al. [114] investigated training strategies and DNNs to improve the retrieval performance. It is proved that better training strategies, better data augmentation, and better structural refinement could achieve better FIR results. As is shown in Table 2, we make a comparison of some methods on two subsets of DeepFashion dataset.

5.2. Person Re-IDentification

Person Re-IDentification (Re-ID), also known as Person Retrieval, is to match the images of the same individual captured on different camera views and is usually considered as a sub-problem of image retrieval. Person Re-ID can retrieve specific pedestrian targets in the cross-device image video and make up for the viewing angle limitations of the current fixed camera. It has wide applications, such as intelligent security, intelligent video surveillance, intelligent retrieval, etc. There are three main challenges in person Re-ID. (1) The images from different camera views differ from each other significantly because of the variation of background and appearance (e.g., illumination, pose, occlusion, resolution). (2) There exists some interference of similar images with different identities. (3) Changes in human pose and human body occlusion may make the problem more complicated. To address these challenges, numerous efforts have been paid from different theoretical perspectives.

A branch of these works is to learn effective representations for improving retrieval performance. In this branch, some person Re-ID methods are based on attention [115,116]. Mancs [115] solved the person re-identification problem by fully utilizing the attention mechanism for the person misalignment problem. Li et al. [116] showed the advantages of jointly learning attention selection and feature representation in a CNN by maximizing the complementary information of different levels of visual attention subject to Re-ID discriminative learning constraints. Inspired by effective human posture estimation, some person Re-ID methods are guided by pose [117–119]. Li et al. [117] proposed a novel method to learn and localize deformable pedestrian parts using Spatial Transformer Networks (STN) with novel spatial constraints. Saquib et al. [118] proposed an effective approach that incorporated both the fine and coarse pose information of the person to learn a discriminative embedding. Furthermore, Attention-aware Feature Composition (AFC) [119] estimated pose-guided visibility scores for body parts to deal with part occlusion in the proposed AFC module. Some person Re-ID methods are guided by mask because neglecting the problem of background clutter can lead to degraded performance [120]. In order to alleviate the problem of cluttered background,

Table 1
The list of the most commonly used datasets in fashion image retrieval.

Dataset	Images	Categories	pairs	year
DARN [101]	182,780	20	91,390	2015
WTBI [102]	78,958	11	39479	2015
DeepFashion [103]	800,000	50	251,000	2016
Modanet [104]	55,000	13	-	2018
DeepFashion2 [105]	491,000	13	873,000	2019

Table 2
A comparison of some methods on DeepFashion dataset. The evaluation metric is top-k recall (R@k). C2S and IS indicate the consumer-to-shop subset of DeepFashion and the in-shop subset of DeepFasion respectively.

Method	DeepFashion(C2S)				DeepFashion(IS)			
	R@1	R@5	R@20	R@50	R@1	R@5	R@20	R@50
WTBI [102]	0.024	0.035	0.063	0.087	0.347	0.424	0.506	0.541
DARN [101]	0.036	0.063	0.111	0.152	0.381	0.547	0.675	0.716
FashionNet [103]	0.073	0..121	0.188	0.226	0.532	0.678	0.764	0.800
VAM [107]	0.128	0.280	0.431	0.568	0.666	–	0.923	–
Gajic et al. [106]	–	0.250	0.450	–	–	–	–	–
Minchul et al. [113]	**0.265**	**0.497**	**0.664**	**0.755**	**0.887**	**0.961**	**0.984**	**0.991**

Qi et al. [120] incorporated masked images with only the foreground regions and input them to the proposed neural network. Some person RE-ID methods employ GAN [121–123]. Martinel et al. [121] applied a dictionary-learning scheme to transfer the feature learned by object recognition and person detection to target re-identification domain. Wei et al. [122] proposed a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap which essentially caused severe performance drop when training and testing on different datasets. Camera Style (CamStyle) [123] used a GAN serve as a data augmentation approach that reduces the risk of overfitting and smooths the CamStyle disparities.

Another branch of these works pays efforts in deep metric learning. Most existing Re-ID frameworks are optimized by contrastive loss or triplet loss [124] or quadruplet loss [59]. Cheng et al. [124] introduced a pull term into the triplet loss to penalize large distances between positive embeddings. Chen et al. [59] added another pull term for the distance between negative pairs in quadruplet loss, which could get a model with a larger inter-class variation and a smaller intra-class variation. Bai et al. [125] concentrated on re-ranking with the capacity of metric fusion for Re-ID. The proposed Unified Ensemble Diffusion (UED) is an effective algorithm which achieves the state-of-the-art retrieval performance on Market-1501 dataset. In addition to the above algorithms for improving search performance, some researchers tried to build a baseline for person Re-ID [126,127]. A good baseline is very important to methods comparison and evaluation. Firstly, it essentially leads to an unfair comparison between the different person Re-ID approaches without a public baseline. Secondly, it is significant to judge the objective capacity of the existing CNN-based person re-identification methods. To deal with the above problems, Xiong et al. [126] proposed three practices for building an effective CNN baseline model towards person reidentification: batch normalization after the global pooling layer, executing identity categorization directly using only one fully-connected layer, and using Adam as optimizer. Furthermore, Luo et al. [127] collected and evaluated some effective training tricks in person Re-ID, then combined these tricks to build a well-performed baseline which achieves high performance.

With the extensive application of deep learning in person Re-ID, several large-scale datasets have been published. They are given in Table 3. We compare some Re-ID methods on Market1501 dataset and DukeMTMC dataset. And, the results are shown in Table 4.

5.3. E-commerce product retrieval

E-commerce product retrieval is one of the most important parts on an E-commerce Platform such as Alibaba [131], JD [132], eBay [133] and Walmart [134]. In e-commerce shopping, consumers usually do not know the correct keywords used to find their desired items. E-commerce product retrieval can help consumers search the products they want. However, this task is challenging. First, it is very difficult to handle heterogeneous image data and bridge the gap between real-shot images from users and the online images. Second, dealing with large scale indexing for massive updating data is also not an easy thing.

Recently, some researchers have made great efforts to this problem. Since the background of product images is significantly irrelevant to the product, Wang et al. [135] utilized the saliency box to filter the proposals extracted by selective searching, then proposed Channel Weighting Generalized Mean Pooling (CWGMP) feature which preserved the discriminability and correlation of convolution features to improve retrieval performance. Pailitao in Alibaba [131] focused on building a real-time and stable search engine. It uses binary indexing engine and re-ranking to improve the engagements, which allow users to freely take photos to find identical items with millisecond response and lossless recall in a highly available and scalable solution. However, Pailitao does not effectively handle database update issues. For this, JD [132] handled frequent image updates through distributed hierarchical architecture and efficient indexing methods. Although implicit feedback, such as page views and click logs, allows for model training with a triplet loss even in Alibaba [131], implicit feedback is available only after launching a visual search system into production. Yamaguchi et al. [136] proposed an image representation method with query feature transformation which narrows the gap between query vector and image vectors on retrieval dataset. In addition, Magnani et al. [134] did an exploration of trained various product retrieval models trained on search log data to further improve retrieval performance. All of the above algorithms have their own datasets, which are not public due to commercial competition and other reasons.

5.4. Remote sensing image retrieval

With the development of remote sensing technologies, the quantity and quality of remote sensing images have increased dramatically. Remote sensing images can be used in some fields to

Table 3
The list of the most commonly used datasets in person re-identification.

Dataset	Images	Identities	Cameras	year
CUHK03 [128]	12,697	1,467	5	2014
Market1501 [129]	32,668	1,501	6	2015
DukeMTMC [130]	34,183	1,402	8	2016
MSMT17 [122]	114,782	4,101	15	2018

Table 4
A comparison of some Re-ID methods on Market1501 dataset and DukeMTMC dataset. The performance is measured via rank-1 accuracy (r = 1) and mean Average Precision (mAP).

Type		Market1501		DukeMTMC	
		$r = 1$	mAP	$r = 1$	mAP
Attention-based	Mancs [115]	93.1	82.3	84.9	71.8
	Li et al. [116]	91.2	75.7	80.5	63.8
Pose-guided	AFC [119]	85.9	66.8	76.8	59.3
	Saquib et al. [118]	78.7	56.0	–	–
Mask-guided	Qi et al. [120]	90.0	75.3	78.8	61.9
Gan-based	CamStyle [123]	88.1	68.7	75.3	53.5
Global feaure	Xiong et al. [126]	92.5	79.8	83.5	68.5
	Luo et al. [127]	94.5	85.9	86.4	76.4
	UED [125]	**95.9**	**92.8**	–	–

solve important problems, such as weather prediction, climate monitoring, urban planning, geological analysis, disaster monitoring, resource investigation, and so on. Among them, Content-Based Remote Sensing Image Retrieval (CBRSIR) is a key problem that could effectively use these remote sensing data. It can automatically and efficiently retrieve the remote sensing images that users need from large scale image databases and has attracted extensive attention from researchers all over the world. According to the improvement of the method, we can divide the CBRSIR methods in the past decade into two categories: feature-based CBRSIR methods and hash-based CBRSIR methods.

Feature-based approaches improve retrieval performance by extracting more discriminative and powerful features. Early methods mainly used conventional features [137,138]. Shao et al. [137] combined color and texture features to improve the performance of RSIR. Color-Texture-Structure-Spectral Speeded Up Robust Features (CTSS-SURF) [138] is a novel local representation for remote sensing image. It is achieved by dividing images into several parts and then designing regional feature vectors. In this way, it can effectively overcome the challenges of RSIR, such as scale, illumination, shift, and rotation variation. Some methods begin to use CNN features through deep learning techniques [139–142]. Li et al. [139] combined the deep features and conventional features to represent remote sensing images, then use collaborative affinity metric fusion to get retrieval results. Zhou et al. [140] proposed two effective schemes for RSIR: fine-tuning the pre-trained CNNs on a remote sensing dataset and using a novel CNN architecture based on convolutional layers and a three-layer perceptron which has fewer parameters to learn low dimensional features from limited labeled images. Hu et al. [143] introduced multiscale concatenation for convolutional features and multipatch pooling for fully connected layers to RSIR. Some methods employ deep metric learning to extract more discriminative features for RSIR [141]. Cao et al. [141] presented a novel triplet deep neural network-based metric learning method to enhance RSIR. Using this method, they embedded the remote sensing images into a semantic space in which images from the same class were close to each other and those from different classes were distinguishable from each other. Some methods adopt deep attention mechanisms to improve retrieval performance [142,144]. Xiong et al. [142] proposed a new attention module for feature extraction for CBRSIR, which could pay more attention to the salient features, and suppress the less useful ones. This attention module can be easily embedded with the last convolutional layer of any pre-trained CNNs and can be applied along two dimensions: channel and spatial axises, attending to emphasize the meaningful features along these two axes. Ye et al. [144] presented an RSIR method based on weighted distance and CNN. First, it uses the fine-tuned CNN models to extract image features and label the class of image in the retrieval dataset. Then it calculates the weight of each class according to the probability of the query image belonging to each class, and uses it to adjust the distance between the query image and the retrieved images.

Hash-based CBRSIR methods speed up retrieval in large-scale dataset by optimizing database search [145–147]. Demir et al. [145] introduced a kernel-based hashing method. It learns hash functions in the kernel space from handcrafted features to enhance the retrieval efficiency. Li et al. [146] introduced a Deep Hashing Neural Network (DHNN), which could jointly learn the deep features and deep hashing code utilizing a cross-entropy loss, for large-scale RSIR. However, the absence of a margin threshold between positive and negative samples in DHNN may lead to poor generalization. To address this problem, Roy et al. [147] presented Metric-Learning based deep hashing Network (MiLaN) that learned a semantic-based metric space, while simultaneously producing binary hash codes for fast and accurate retrieval of RS images.

The most commonly used datasets in RSIR are shown in Table 5. We also compare some CBRSIR methods on UC-Merced dataset. And, the results are shown in Table 6.

5.5. Trademark image retrieval

Trademark is the symbol of enterprise brand. Protecting the trademarks from infringements and piracy is of great significance to promote the development of enterprises and protect the interests of consumers. Trademark Image Retrieval (TIR) can search all trademark images that are similar or related to a given input from a trademark dataset. Trademark images are manually designed artificially and very different from natural scene images. They usually consist of graphical and textual primitives, where the color and typeface can be changeable and artistic. So some natural scene images features are not powerful enough for describing complicated trademark images. Various approaches have been proposed for trademark image retrieval.

Some TIR methods use conventional features [153,154]. Tursun et al. [153] proposed the color histogram, gradient orientation histogram, LBP, shape context, SIFT and triangular SIFT features to search trademark images on their published dataset. Feng et al. [154] extracted reversal invariant SIFT features from edges of the segmented blocks of a trademark, then aggregated SIFT features from each block to generate a single global representation. These methods have limited performance owing to limited representation and complex calculations of conventional features. Some methods use CNN features [155–158]. Aker et al. [155] provided analysis on TR with deep features, and showed that deep features were superior to conventional features. Lan et al. [156] utilized mid-level convolutional features extracted from a pre-trained network and applied uniform Local Binary Patterns (LBP) to features maps for aggregation. Tursun et al. [157] provided a large-scale dataset with benchmark queries, METU dataset. And, they pro-

Table 5
The list of the most commonly used datasets in remote sensing image retrieval.

Dataset	Images	Categories	Resolution (m)	Year
UC-Merced [148]	2,100	21	0.3	2010
AID [149]	10,000	30	0.5–0.8	2017
NWPU-RESISC45 [150]	31,500	45	0.2–30	2017
PatternNet [151]	30,400	38	0.062–4.693	2018
AID++ [152]	400,000	46	–	2018

Table 6
A comparison of some CBRSIR methods on UC-Merced dataset. The performance is measured via Average Normalized Modified Retrieval Rank(ANMRR), mean Average Precision (mAP) and precision at k (P@k). For ANMRR, lower values indicate better performance, while for mAP and P@k, larger is better.

Method	ANMRR	mAP	p@5	p@10	p@100
CTSS-SURF [138]	0.1470	0.8124	–	0.9680	0.7502
Zhou et al. [140]	0.3750	–	–	–	–
Hu et al. [143]	0.2850	–	–	–	–
Ye et al. [144]	0.0404	–	–	–	–
Xiong et al. [142]	0.0890	0.8400	0.9190	0.9140	–
Cao et al. [141]	**0.0223**	**0.9663**	**0.9775**	**0.9757**	**0.4855**

vided a baseline on this benchmark using the widely-used methods applied to TIR in the literature. Recently, they proposed both hard and soft attention approaches [158], which directly focus on critical information and reduce the attention given to distracting and uninformative elements. This method achieved a new state-of-the-art result on the METU dataset.

There are several wildly used datasets for trademark retrieval, such as METU dataset [153], FlickrLogos [159] and NPU-TM [156]. These datasets are shown in the Table 7. We further compare some TIR methods on METU dataset with the experimental results shown in Table 8.

Nowadays, wine culture has gradually integrated into our daily life. As a result, wine label image retrieval has become an important and emergent task. Although wine label can be considered as a special trademark, the task of wine label image retrieval is quite different from the general trademark image retrieval on two aspects: (1). the input images of wine label image retrieval systems are usually taken by a mobile phone without preprocessing. These kinds of images contain complex backgrounds. In contrast, the general trademark image retrieval aims only at processing noiseless trademark images, which is separated from backgrounds. (2). The goal of the general trademark image retrieval is just to return the trademark of each input image, however, a wine label image retrieval system needs to output not only main-brand (trademark) but also sub-brand (more fine-grained information, such as production year and location), to help a consumer purchase the wine or learn more about the wine taken by him. In fact, wine label image retrieval has two major challenges. First, there is a huge number of wine label images with a large number of brands. Moreover, the numbers of samples in different brands are various; Second, there is a significant difference among many wine label images of the same brand, while the difference among some wine label images of different brands is not obvious. These challenges make the wine label image retrieval rather difficult.

In the last decade, some researchers have made some efforts for wine label image retrieval. Lim et al. [160] searched for wine label images by recognizing the fonts of the brand texts. They firstly got wine label regions by using an edge-based method, and then used fuzzy c-means clustering in local regions of individual wine characters to recognize these texts. While this system can achieve a high recognition accuracy in some special wine label images, it has some disadvantages obviously. First, it works well only if the texts on the wine label are English. Second, the character recognition in this system relies heavily on the detection of candidate text regions. However, the detection of candidate text regions by the edge-based method is usually inaccurate in where the font style is changeable and the sizes of characters are quite various. To improve the detection accuracy of candidate text regions, Wang et al. [161] proposed a new local Chan-Vese (LCV) model, which is based on the techniques of curve evolution, local statistical function and level set method. Particularly, when the LCV model was combined with an extended structure tensor by adding the intensity information into the classical structure tensor for texture image segmentation, the texture image can be efficiently segmented no matter whether it presents intensity in homogeneity or not. In fact, the LCV model can be efficient for the two-modal (phase) images, which usually generate two segments, i.e., foreground and background. As a bimodal model, it cannot simultaneously detect multiple objects in different intensities. The texts of wine label images are usually in different intensities because of artistic designs of the text fonts. So, the LSV method may not be the most effective for detecting text. Wu et al. [162] further used certain hierarchical features and a client-server architecture to search the image from the retrieval dataset. Particularly they utilized SURF descriptors, K-D tree and *k*-means methods to build the retrieval system. However, there still exist certain disadvantages in this mechanism. Leaving aside the fact that the SURF descriptor is a kind of conventional local feature which can not reduce the semantic gap in retrieval, the retrieval time of each image will increase rapidly when the retrieval is implemented on a large dataset. To address these problems, Li et al. proposed the CNN-SIFT Consecutive Searching and Matching (CSCSM) framework [163] and CNN-SURF Consecutive Filtering and Matching

Table 7
The list of the most commonly used datasets in trademark image retrieval

Dataset	Images	Categories	Background	Year
FlickrLogos [159]	8240	32	Yes	2011
METU dataset [153]	930,328	409,834	No	2015
NPU-TM [156]	7139	317	No	2017

Table 8
A comparison of some TIR methods on METU dataset on the Normalized Average Ranks (NARs) where the smaller NAR indicates the better result.

Type	Method	NAR
Conventional feature-based	Feng et al. [154]	0.083
	TRI-SIFT [157]	0.324
	Surf [157]	0.207
	OR-SIFT [157]	0.190
Deep feature-based	VggNet [157]	0.086
	GoogLeNet [157]	0.118
	Tursun et al. [157]	0.062
	Tursun et al. [158]	**0.040**

(CSCFM) framework [164] for wine label retrieval with a large number of brands. The CSCSM framework firstly utilized an advanced deep CNN to shrink the search range by recognizing the main-brand in a supervised learning mode, and then applied an improved SIFT descriptor based on the combination of the RANdom SAmple Consensus (RANSAC) and Term Frequency-Inverse Document Frequency (TF-IDF) mechanisms to match the final sub-brand. The CSCFM framework improved and extended the previous study of the CSCSM framework methodologically and theoretically. It utilized a new version of CNN architecture and an improved SURF matching scheme by adopting the RANSAC and modified TF-IDF distance that can reduce the computational cost and improve the retrieval performance greatly. Both the CSCSM and CSCFM frameworks can not only retrieve the main-brand but also find out the sub-brand. Moreover, they can implement the wine label retrieval, and they all can increase the retrieval accuracy on a large dataset effectively and efficiently.

6. Future research directions

By the above survey, it is clear that CBIR has made great progress on theory, technology and application in the past decade. However, there are still many challenges, especially with the emergence of big data and the utilization of deep learning techniques. In this section, we discuss these challenges and give some potential research directions of CBIR in the future.

6.1. Collecting more and larger datasets

One critical direction of CBIR in future research is to collect more and larger datasets. Deep learning techniques are data-driven. In general, as long as there emerge new and large scale datasets, we can train the good deep neural network models to refresh the retrieval accuracy and solve the database search problems. However, in the training process, the over-fitting problems may hinder the breakthrough of the learning algorithm. So, more and larger datasets are necessary and valuable. For the general instance retrieval, more and larger instance datasets can make the search applicable to many search purposes. If the CNN in a CBIR system is trained with a larger dataset which combines the large person re-identification dataset and large e-commerce product retrieval dataset, it may be able to efficiently apply to both clothing search and commodity search. For various specialized instance retrieval, more and larger datasets are also crucial to the performance of retrieval. The new state-of-art methods can be only established with the larger scale and richer forms of datasets. To effectively use the dataset, the label of the new dataset should be accurate enough to eliminate some ambiguity problems in the relevance of image content, such as commodity icon data.

6.2. Establishing the effective learning strategies for small scale image retrieval

Another critical direction of CBIR in future research is to establish some effective learning strategies for small scale image retrieval. At present, most CBIR methods require large enough, even massive datasets. As for instance retrieval tasks, the cost of image collection and labeling is difficult and expensive, which limits the development of CBIR in real-world scenarios. However, human beings have the ability to learn new concepts with little supervision information. For example, an adult can find the most similar images without supervision. In order to enable CBIR to have the same learning ability through a small number of training samples as human beings, the researchers need to establish more effective learning strategies with a small scale training dataset so that the CBIR methods can make the large scale database search in various real-world scenarios.

6.3. Establishing more efficient database search

In practical applications of instance retrieval, especially where the reference dataset is extremely large, the time cost of searching the nearest neighbor of an input image is awfully expensive. Due to the advantages of easy implementation, fast query speed and low storage cost, Hashing, especially deep Hashing, has been wildly deployed to retrieval tasks on large-scale datasets. However, Hashing based methods face another challenge of losing precision when transforming feature vectors into binary encoding, which will induce a sub-optimal retrieval result and further decrease the searching accuracy. Therefore, how to establish a more efficient and precise database search is our major breakthrough of the future work.

6.4. Adoption of the automated machine learning and neural architecture search

Currently, the most state-of-the-art architectures of deep neural networks are designed artificially, however, this has been becoming a limitation considering the rapidly developed computation ability of machines and the “laggard” human knowledge. In order to solve this problem, some new deep learning neural network can design their own best architectures to a learning task. Automated Machine Learning (AutoML) and Neural Architecture Search (NAS) are two such models attracting much attention in the computer vision community. Actually, many recent methods based on AutoML or NAS have already achieved the state-of-the-art results and outperformed the and-designed architectures in various computer vision applications. Therefore, it is worth trying to apply both AutoML and NAS into the task of instance retrieval to get more effective image representation.

7. Conclusion

In this survey, we have summarized recent developments of Content-Based Image Retrieval from technological and practical applications in the last decade. First, we review the developments of image representation (or feature extraction) and database search for CBIR. We then present the typical practical applications of CBIR on fashion image retrieval, person re-identification, e-commerce product retrieval, remote sensing image retrieval and trademark label image retrieval, respectively. Finally, we discuss the challenges and potential research directions in the future with the emergence of big data and the utilization of deep learning techniques. It is clear that CBIR has developed into a new era and will

play an important role in artificial intelligence in the future for our daily life.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key R & D Program of China (2018YFC0808305).

References

[1] Z. Huang, R. Wang, S. Shan, X. Chen, Projection metric learning on grassmann manifold with application to video based face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 140–149.
[2] Y. Zhang, P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, R. Jin, Visual search at alibaba, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 993–1001.
[3] A. Jimenez, J.M. Alvarez, X. Giro-i Nieto, Class-weighted convolutional features for visual instance search, arXiv preprint arXiv:1707.02581.
[4] X. Han, Z. Wu, Y.-G. Jiang, L.S. Davis, Learning fashion compatibility with bidirectional lstms, in: Proceedings of the 25th ACM international conference on Multimedia ACM, 2017, pp. 1078–1086.
[5] M. Yasmin, M. Sharif, S. Mohsin, Neural networks in medical imaging applications: a survey, World Appl. Sci. J. 22 (1) (2013) 85–96.
[6] A. Alzu'Bi, A. Amira, N. Ramzan, Semantic content-based image retrieval: a comprehensive study, J. Vis. Commun. Image Represent. 32 (2015) (2015) 20–54.
[7] L. Zheng, Y. Yang, Q. Tian, Sift meets cnn: a decade survey of instance retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1224–1244.
[8] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Revisiting oxford and paris: Large-scale image retrieval benchmarking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5706–5715.
[9] W. Zhou, H. Li, Q. Tian, Recent advance in content-based image retrieval: A literature survey, arXiv preprint arXiv:1706.06064.
[10] J. Wang, X. Hua, Interactive image search by color map, ACM Trans. Intell. Syst. Technol. 3(1) (2011) 12.
[11] S. Bai, X. Bai, Z. Zhou, Z. Zhang, L. Jan Latecki, Gift A real-time and scalable 3d shape search engine, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5023–5032.
[12] X. Wang, B. Zhang, H. Yang, Content-based image retrieval by integrating color and texture features, Multimedia Tools Appl. 68 (3) (2014) 545–569.
[13] Y. Li, L. Shapiro, J.A. Bilmes, A generative/discriminative learning algorithm for image classification, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 2, IEEE, 2005, pp. 1605–1612.
[14] A.P. Berman, L.G. Shapiro, A flexible image database system for content-based retrieval, Comput. Vis. Image Underst. 75 (1–2) (1999) 175–195.
[15] Z.-Q. Zhao, H. Glotin, Z. Xie, J. Gao, X. Wu, Cooperative sparse representation in two opposite directions for semi-supervised image annotation, IEEE Trans. Image Process. 21 (9) (2012) 4218–4231.
[16] W. Zhou, Y. Lu, H. Li, Q. Tian, Scalar quantization for large scale image search, in: Proceedings of the 20th ACM international conference on Multimedia ACM, 2012, pp. 169–178.
[17] W. Zhou, H. Li, R. Hong, Y. Lu, Q. Tian, Bsift: toward data-independent codebook for large scale image search, IEEE Trans. Image Process. 24 (3) (2015) 967–979.
[18] S. Zhang, Q. Tian, K. Lu, Q. Huang, W. Gao, Edge-sift: discriminative binary descriptor for scalable partial-duplicate mobile search, IEEE Trans. Image Process. 22 (7) (2013) 2889–2902.
[19] R. Sicre, T. Gevers, Dense sampling of features for image retrieval, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 3057–3061.
[20] S. Madeo, M. Bober, Fast, compact and discriminative: evaluation of binary descriptors for mobile applications, IEEE Trans. Multimedia PP 99 (2016) 1–1.
[21] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Null, IEEE, 2003, p. 1470.
[22] F. Perronnin, Y. Liu, J. Sánchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3384–3391.
[23] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3304–3311.
[24] H. Jégou, A. Zisserman, Triangulation embedding and democratic aggregation for image search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3310–3317.
[25] Z. Gao, J. Xue, W. Zhou, S. Pang, Q. Tian, Fast democratic aggregation and query fusion for image search, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 35–42.
[26] L. Zhen, H. Li, W. Zhou, R. Hong, T. Qi, Uniting keypoints: local visual information fusion for large scale image search, IEEE Trans. Multimedia 17 (4) (2015) 538–548.
[27] R. Arandjelovic, A. Zisserman, All about vlad, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2013, pp. 1578–1585.
[28] E. Spyromitros-Xioufis, S. Papadopoulos, I.Y. Kompatsiaris, G. Tsoumakas, I. Vlahavas, A comprehensive study over vlad and product quantization in large-scale image retrieval, IEEE Trans. Multimedia 16 (6) (2014) 1713–1728.
[29] S.S. Husain, M. Bober, Improving large-scale image retrieval through robust aggregation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2017) 1783–1796.
[30] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural codes for image retrieval (2014) 584–599.
[31] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition (2014) 806–813.
[32] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features (2014) 392–407.
[33] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, C. Schmid, Local convolutional features with unsupervised training for image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 91–99.
[34] M. Teichmann, A. Araujo, M. Zhu, J. Sim, Detect-to-retrieve, Efficient regional aggregation for image search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5109–5118.
[35] K. Reddy Mopuri, R. Venkatesh Babu, Object level deep feature pooling for compact image representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 62–70.
[36] K. Yan, Y. Wang, D. Liang, T. Huang, Y. Tian, Cnn vs. sift for image retrieval: alternative or complementary?, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp 407–411.
[37] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of cnn activations, arXiv preprint arXiv:1511.05879.
[38] A. Babenko, V. Lempitsky, Aggregating local deep features for image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1269–1277.
[39] A. Iscen, Y. Avrithis, G. Tolias, T. Furon, O. Chum, Fast spectral ranking for similarity search (2018) 7632–7641.
[40] X. Wei, J. Luo, J. Wu, Z. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, IEEE Trans. Image Process. 26 (6) (2017) 2868–2881.
[41] Y. Kalantidis, C. Mellina, S. Osindero, Cross-dimensional weighting for aggregated deep convolutional features, in: European Conference on Computer Vision, Springer, 2016, pp. 685–701.
[42] A. Jimenez, J.M. Alvarez, X. Giro-i Nieto, Class-weighted convolutional features for visual instance search, arXiv preprint arXiv:1707.02581.
[43] J. Xu, C. Shi, C. Qi, C. Wang, B. Xiao, Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 1–8.
[44] H. Noh, A. Araujo, J. Sim, T. Weyand, B. Han, Large-scale image retrieval with attentive deep local features, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3456–3465.
[45] J. Kim, S. Yoon, Regional attention based deep feature for image retrieval, in: BMVC, 2018, p. 209.
[46] L. Yang, Y. Xu, J. Wang, M. Zhuang, Y. Zhang, Ms-rmac: multi-scale regional maximum activation of convolutions for image retrieval, IEEE Signal Process. Lett. 99 (2017) 1–1.
[47] O. Seddati, S. Dupont, S. Mahmoudi, M. Parian, Towards good practices for image retrieval based on cnn features, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1246–1255.
[48] S.S. Husain, M. Bober, Remap: multi-layer entropy-guided pooling of dense cnn features for image retrieval, IEEE Trans. Image Process. 28 (10) (2019) 5201–5213.
[49] A. Gordo, J. Almazn, J. Revaud, D. Larlus, End-to-end learning of deep visual representations for image retrieval, Int. J. Comput. Vision (2017) 1–18.
[50] F. Radenović, G. Tolias, O. Chum, Fine-tuning cnn image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1655–1668.
[51] H. Ohsong, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012.
[52] H. Oh Song, S. Jegelka, V. Rathod, K. Murphy, Deep metric learning via facility location, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5382–5390.
[53] J. Revaud, J. Almazan, R.S. de Rezende, C.R. de Souza, Learning with average precision: Training image retrieval with a listwise loss, arXiv preprint arXiv:1906.07589.
[54] S. Kim, M. Seo, I. Laptev, M. Cho, S. Kwak, Deep metric learning beyond binary supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2288–2297.
[55] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.
[56] F. Radenović, G. Tolias, O. Chum, Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples, in: European Conference on Computer Vision, Springer, 2016, pp. 3–20.
[57] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, 2016, pp. 1857–1865.
[58] X. He, Y. Zhou, Z. Zhou, S. Bai, X. Bai, Triplet-center loss for multi-view 3d object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1945–1954.
[59] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[60] K. He, Y. Lu, S. Sclaroff, Local descriptors optimized for average precision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 596–605.
[61] D. Zhang, M.M. Islam, G. Lu, J. Hou, Semantic image retrieval using region based inverted file, in: 2009 Digital Image Computing: Techniques and Applications, IEEE, 2009, pp. 242–249.
[62] J. Cai, Q. Liu, F. Chen, D. Joshi, Q. Tian, Scalable image search with multiple index tables, in: Proceedings of International Conference on Multimedia Retrieval, ACM, 2014, p. 407.
[63] I. Bartolini, M. Patella, Windsurf: the best way to surf, Multimedia Syst. 24 (4) (2018) 459–476.
[64] A. Babenko, V. Lempitsky, The inverted multi-index, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2014) 1247–1260.
[65] L. Zheng, S. Wang, Q. Tian, Coupled binary embedding for large-scale image retrieval, IEEE Trans. Image Process. 23 (8) (2014) 3368–3380.
[66] Z. Liu, H. Li, L. Zhang, W. Zhou, Q. Tian, Cross-indexing of binary sift codes for large-scale image search, IEEE Trans. Image Process. 23 (5) (2014) 2047–2057.
[67] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, H. Winnemoeller, Recognizing image style, arXiv preprint arXiv:1311.3715.
[68] S. Zhang, M. Yang, X. Wang, Y. Lin, Q. Tian, Semantic-aware co-indexing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 37(12) (2015) 2573–2587.
[69] L. Zheng, S. Wang, Z. Liu, Q. Tian, Fast image retrieval: Query pruning and early termination, IEEE Trans. Multimedia 17 (5) (2015) 648–659.
[70] R. Ji, L. Duan, J. Chen, L. Xie, H. Yao, W. Gao, Learning to distribute vocabulary indexing for scalable visual search, IEEE Trans. Multimedia 15 (1) (2013) 153–166.
[71] Y. Xia, K. He, F. Wen, J. Sun, Joint inverted indexing, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3416–3423.
[72] A. Babenko, V. Lempitsky, Efficient indexing of billion-scale datasets of deep descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2055–2063.
[73] M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Proceedings of the Twentieth Annual Symposium on Computational Geometry, ACM, 2004, pp. 253–262.
[74] W. Kong, W.-J. Li, Isotropic hashing, in: Advances in Neural Information Processing Systems, 2012, pp. 1646–1654.
[75] H. Liu, R. Ji, Y. Wu, W. Liu, Towards optimal binary code learning via ordinal embedding, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016, p. 674–685.
[76] A. Alzu'bi, A. Amira, N. Ramzan, Content-based image retrieval with compact deep convolutional features, Neurocomputing 249 (2017) 95–105.
[77] X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, L. Yang, Pairwise based deep ranking hashing for histopathology image classification and retrieval, Pattern Recogn. 81 (2018) 14–22.
[78] S. Cheng, L. Wang, A. Du, An adaptive and asymmetric residual hash for fast image retrieval, IEEE Access 7 (2019) 78942–78953.
[79] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search., in: ICCV, vol. 9, 2009, pp. 2130–2137.
[80] Q. Jiang, W. Li, Scalable graph hashing with feature transformation, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 2248–2254.
[81] G. Lin, C. Shen, Q. Shi, A. Van den Hengel, D. Suter, Fast supervised hashing with decision trees for high-dimensional data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1963–1970.
[82] F. Shen, C. Shen, W. Liu, H. Tao Shen, Supervised discrete hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 37–45.
[83] W. Kang, W. Li, Z. Zhou, Column sampling based discrete supervised hashing, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 1230–1236.
[84] F. Shen, Y. Yang, L. Liu, W. Liu, D. Tao, H.T. Shen, Asymmetric binary coding for image search, IEEE Trans. Multimedia 19 (9) (2017) 2022–2032.
[85] C. Ma, I.W. Tsang, F. Peng, C. Liu, Partial hash update via hamming subspace learning, IEEE Trans. Image Process. 26 (4) (2017) 1939–1951.
[86] R. Xia, Y. Pan, H. Lai, C. Liu, S. Yan, Supervised hashing for image retrieval via image representation learning, in: Twenty-eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2156–2162.
[87] H. Lai, Y. Pan, Y. Liu, S. Yan, Simultaneous feature learning and hash coding with deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3270–3278.
[88] S. Conjeti, A.G. Roy, A. Katouzian, N. Navab, Hashing with residual networks for image retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 541–549.
[89] F. Zhao, Y. Huang, L. Wang, T. Tan, Deep semantic ranking based hashing for multi-label image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1556–1564.
[90] H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2064–2072.
[91] C. Zhou, L.-M. Po, M. Liu, W.Y. Yuen, P.H. Wong, H.-T. Luk, K.W. Lau, H.K. Cheung, Deep hashing with triplet labels and unification binary code selection for fast image retrieval, in: International Conference on Multimedia Modeling, Springer, 2019, pp. 277–288.
[92] Y. Li, L. Wan, T. Fu, W. Hu, Piecewise supervised deep hashing for image retrieval, Multimedia Tools Appl. (2019) 1–21.
[93] Y. Cao, M. Long, J. Wang, S. Liu, Deep visual-semantic quantization for efficient image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1328–1337.
[94] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2010) 117–128.
[95] T. Ge, K. He, Q. Ke, J. Sun, Optimized product quantization for approximate nearest neighbor search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2946–2953.
[96] T. Zhang, G.-J. Qi, J. Tang, J. Wang, Sparse composite quantization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4548–4556.
[97] Y. Cao, M. Long, J. Wang, H. Zhu, Q. Wen, Deep quantization network for efficient image retrieval, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 3457–3463.
[98] T. Yu, J. Yuan, C. Fang, H. Jin, Product quantization network for fast image retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 186–201.
[99] J. Song, X. Zhu, L. Gao, X.-S. Xu, W. Liu, H.T. Shen, Deep recurrent quantization for generating sequential binary codes, arXiv preprint arXiv:1906.06699.
[100] L. Gao, X. Zhu, J. Song, Z. Zhao, H.T. Shen, Beyond product quantization: deep progressive quantization for image retrieval, arXiv preprint arXiv:1906.06698.
[101] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1062–1070.
[102] M. Hadi Kiapour, X. Han, S. Lazebnik, A.C. Berg, T.L. Berg, Where to buy it: Matching street clothing photos in online shops, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3343–3351.
[103] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[104] S. Zheng, F. Yang, M.H. Kiapour, R. Piramuthu, Modanet, A large-scale street fashion dataset with polygon annotations, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 1670–1678.
[105] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, Deepfashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5337–5345.
[106] B. Gajic, R. Baldrich, Cross-domain fashion image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1869–1871.
[107] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, X. Gu, Clothing retrieval with visual attention model, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.
[108] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, X.-S. Hua, An adversarial approach to hard triplet generation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 501–517.
[109] Y. Yuan, K. Yang, C. Zhang, Hard-aware deeply cascaded embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 814–823.
[110] A. Chopra, A. Sinha, H. Gupta, M. Sarkar, K. Ayush, B. Krishnamurthy, Powering robust fashion retrieval with information rich feature embeddings, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[111] Q. Dong, S. Gong, X. Zhu, Multi-task curriculum transfer deep learning of clothing attributes, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 520–529.
[112] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, R. Feris, Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5334–5343.
[113] M. Shin, S. Park, T. Kim, Semi-supervised feature-level attribute manipulation for fashion image retrieval, CoRR abs/1907.05007. arXiv:1907.05007. http://arxiv.org/abs/1907.05007.
[114] S. Park, M. Shin, S. Ham, S. Choe, Y. Kang, Study on fashion image retrieval methods for efficient fashion visual search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[115] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: a multi-task attentional network with curriculum sampling for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 365–381.

[116] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.
[117] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.
[118] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 420–429.
[119] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2119–2128.
[120] L. Qi, J. Huo, L. Wang, Y. Shi, Y. Gao, Maskreid: A mask based deep ranking neural network for person re-identification, arXiv preprint arXiv:1804.03864.
[121] N. Martinel, M. Dunnhofer, G.L. Foresti, C. Micheloni, Person re-identification via unsupervised transfer of learned visual representations, in: Proceedings of the 11th International Conference on Distributed Smart Cameras, ACM, 2017, pp. 151–156.
[122] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.
[123] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: a novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019).
[124] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[125] S. Bai, P. Tang, P.H. Torr, L.J. Latecki, Re-ranking via metric fusion for object retrieval and person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 740–749.
[126] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, J.T. Zhou, Good practices on building effective cnn baseline model for person re-identification, in: Tenth International Conference on Graphics and Image Processing (ICGIP 2018), vol. 11069, International Society for Optics and Photonics, 2019, p. 110690l.
[127] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[128] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[129] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[130] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 17–35.
[131] Y. Zhang, P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, R. Jin, Visual search at alibaba, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 993–1001.
[132] J. Li, H. Liu, C. Gui, J. Chen, Z. Ni, N. Wang, Y. Chen, The design and implementation of a real time visual search system on jd e-commerce platform, in: Proceedings of the 19th International Middleware Conference Industry, ACM, 2018, pp. 9–16.
[133] F. Yang, A. Kale, Y. Bubnov, L. Stein, Q. Wang, H. Kiapour, R. Piramuthu, Visual search at ebay, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 2101–2110.
[134] A. Magnani, F. Liu, M. Xie, S. Banerjee, Neural product retrieval at walmart.com, in: Companion Proceedings of The 2019 World Wide Web Conference, ACM, 2019, pp. 367–372.
[135] Z. Wang, X. Liu, H. Li, J. Shi, Y. Rao, A saliency detection based unsupervised commodity object retrieval scheme, IEEE Access 6 (2018) 49902–49912.
[136] T. Yamaguchi, K. Arase, R. Togashi, S. Ueta, Closing the gap between query and database through query feature transformation in c2c e-commerce visual search.
[137] Z. Shao, W. Zhou, Q. Cheng, C. Diao, Z. Lei, An effective hyperspectral image retrieval method using integrated spectral and textural features, Sensor Rev. 35 (3) (2015) 274–281.
[138] S. Bouteldja, A. Kourgli, A.B. Aissa, Efficient local-region approach for high-resolution remote-sensing image retrieval and classification, J. Appl. Remote Sens. 13 (1) (2019) 016512.
[139] Y. Li, Y. Zhang, C. Tao, H. Zhu, Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion, Remote Sens. 8 (9) (2016) 709.
[140] W. Zhou, S. Newsam, C. Li, Z. Shao, Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval, Remote Sens. 9 (5) (2017) 489.
[141] R. Cao, Q. Zhang, J. Zhu, Q. Li, Q. Li, B. Liu, G. Qiu, Enhancing remote sensing image retrieval with triplet deep metric learning network, arXiv preprint arXiv:1902.05818.
[142] W. Xiong, Y. Lv, Y. Cui, X. Zhang, X. Gu, A discriminative feature learning approach for remote sensing image retrieval, Remote Sens. 11 (3) (2019) 281.
[143] F. Hu, X. Tong, G.-S. Xia, L. Zhang, Delving into deep representations for remote sensing image retrieval, in: 2016 IEEE 13th International Conference on Signal Processing (ICSP), IEEE, 2016, pp. 198–203.
[144] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, W. Min, Remote sensing image retrieval using convolutional neural network features and weighted distance, IEEE Geosci. Remote Sens. Lett. 15 (10) (2018) 1535–1539.
[145] B. Demir, L. Bruzzone, Hashing-based scalable remote sensing image search and retrieval in large archives, IEEE Trans. Geosci. Remote Sens. 54 (2) (2015) 892–904.
[146] Y. Li, Y. Zhang, X. Huang, H. Zhu, J. Ma, Large-scale remote sensing image retrieval by deep hashing neural networks, IEEE Trans. Geosci. Remote Sens. 56 (2) (2017) 950–965.
[147] S. Roy, E. Sangineto, B. Demir, N. Sebe, Metric-learning based deep hashing network for content based retrieval of remote sensing images, arXiv preprint arXiv:1904.01258.
[148] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2010, pp. 270–279.
[149] Y. Wang, L. Zhang, H. Deng, J. Lu, H. Huang, L. Zhang, J. Liu, H. Tang, X. Xing, Learning a discriminative distance metric with label consistency for scene classification, IEEE Trans. Geosci. Remote Sens. 55 (8) (2017) 4427–4440.
[150] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proc. IEEE 105 (10) (2017) 1865–1883.
[151] W. Zhou, S. Newsam, C. Li, Z. Shao, Patternnet: a benchmark dataset for performance evaluation of remote sensing image retrieval, ISPRS J. Photogrammetry Remote Sens. 145 (2018) 197–209.
[152] P. Jin, G.-S. Xia, F. Hu, Q. Lu, L. Zhang, Aid++An updated version of aid on scene classification, in: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 4721–4724.
[153] O. Tursun, S. Kalkan, Metu dataset: A big dataset for benchmarking trademark retrieval, in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA), IEEE, 2015, pp. 514–517.
[154] Y. Feng, C. Shi, C. Qi, J. Xu, B. Xiao, C. Wang, Aggregation of reversal invariant features from edge images for large-scale trademark retrieval, in: 2018 4th International Conference on Control, Automation and Robotics (ICCAR), IEEE, 2018, pp. 384–388.
[155] C. Aker, O. Tursun, S. Kalkan, Analyzing deep features for trademark retrieval, in: 2017 25th Signal Processing and Communications Applications Conference (SIU), IEEE, 2017, pp. 1–4.
[156] T. Lan, X. Feng, Z. Xia, S. Pan, J. Peng, Similar trademark image retrieval integrating lbp and convolutional neural network, in: International Conference on Image and Graphics, Springer, 2017, pp. 231–242.
[157] O. Tursun, C. Aker, S. Kalkan, A large-scale dataset and benchmark for similar trademark retrieval, arXiv preprint arXiv:1701.05766.
[158] O. Tursun, S. Denman, S. Sivapalan, S. Sridharan, C. Fookes, S. Mau, Component-based attention for large-scale trademark retrieval, CoRR abs/1811.02746. arXiv:1811.02746. http://arxiv.org/abs/1811.02746.
[159] S. Romberg, L.G. Pueyo, R. Lienhart, R. Van Zwol, Scalable logo recognition in real-world images, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ACM, 2011, p. 25.
[160] J. Lim, S. Kim, J. Park, G. Lee, H. Yang, C. Lee, Recognition of text in wine label images, in: 2009 Chinese Conference on Pattern Recognition, IEEE, 2009, pp. 1–5.
[161] X.-F. Wang, D.-S. Huang, H. Xu, An efficient local chan–vese model for image segmentation, Pattern Recogn. 43 (3) (2010) 603–618.
[162] M. Wu, J. Lee, S. Kuo, A hierarchical feature search method for wine label image recognition, in: 2015 38th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2015, pp. 568–572.
[163] X. Li, J. Yang, J. Ma, Cnn-sift consecutive searching and matching for wine label retrieval, in: International Conference on Intelligent Computing, Springer, 2019, pp. 250–261.
[164] X. Li, J. Yang, J. Ma, Large scale category-structured image retrieval for object identification through supervised learning of cnn and surf-based matching, IEEE Access 8 (2020) 57796–57809.

Xiaoqing Li received the B.S. degree in mathematics from Ocean University of China, Qingdao, China, in 2016, and she is currently pursuing the Ph.D. degree in applied mathematics with the School of Mathematical Sciences, Peking University, Beijing, China. Her current research interests include machine learning, image retrieval and neural networks.

Jiansheng Yang received the B.S., M.S., and Ph.D. degrees from Peking University, Beijing, China, in 1988, 1991, and 1994, respectively. He is currently a Professor of mathematics with Peking University, where he is also a Faculty Member. His research interests include wavelet analysis, image reconstruction, and computer algorithms.

Jinwen Ma received the M.S. degree in applied mathematics from Xi'an Jiaotong University in 1988 and the Ph.D. degree in probability theory and statistics from Nankai University in 1992. From July 1992 to November 1999, he was a lecturer or associate professor at the Department of Mathematics, Shantou University. From December 1999, he became a full professor at the Institute of Mathematics, Shantou University. From September 2001, he has joined the Department of Information Science at the School of Mathematical Sciences, Peking University, where he is currently a full professor and Ph.D. tutor. During 1995 and 2003, he also visited several times at the Department of Computer Science and Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow. He also worked as Research Scientist at Amari Research Unit, RIKEN Brain Science Institute, Japan from September 2005 to August 2006. He has published over 100 academic papers on neural networks, pattern recognition, bioinformatics, and information theory.