Recent Advance in Content-based Image Retrieval: A Literature Survey

Wengang Zhou, Houqiang Li, and Qi Tian *Fellow, IEEE*

Abstract—The explosive increase and ubiquitous accessibility of visual data on the Web have led to the prosperity of research activity in image search or retrieval. With the ignorance of visual content as a ranking clue, methods with text search techniques for visual retrieval may suffer inconsistency between the text words and visual content. Content-based image retrieval (CBIR), which makes use of the representation of visual content to identify relevant images, has attracted sustained attention in recent two decades. Such a problem is challenging due to the intention gap and the semantic gap problems. Numerous techniques have been developed for content-based image retrieval in the last decade. The purpose of this paper is to categorize and evaluate those algorithms proposed during the period of 2003 to 2016. We conclude with several promising directions for future research.

Index Terms—content-based image retrieval, visual representation, indexing, similarity measurement, spatial context, search re-ranking.

1 Introduction

With the universal popularity of digital devices embedded with cameras and the fast development of Internet technology, billions of people are projected to the Web sharing and browsing photos. The ubiquitous access to both digital photos and the Internet sheds bright light on many emerging applications based on image search. Image search aims to retrieve relevant visual documents to a textual or visual query efficiently from a large-scale visual corpus. Although image search has been extensively explored since the early 1990s [1], it still attracts lots of attention from the multimedia and computer vision communities in the past decade, thanks to the attention on scalability challenge and emergence of new techniques. Traditional image search engines usually index multimedia visual data based on the surrounding meta data information around images on the Web, such as titles and tags. Since textual information may be inconsistent with the visual content, content-based image retrieval (CBIR) is preferred and has been witnessed to make great advance in recent years.

In content-based visual retrieval, there are two fundamental challenges, *i.e.*, *intention gap* and *semantic gap*. The intention gap refers to the difficulty that a user suffers to precisely express the expected visual content by a query at hand, such as an example image or a sketch map. The semantic gap originates from the difficulty in describing high-level semantic concept with low-level visual feature [2] [3] [4]. To narrow those gaps, extensive efforts have been made from both the academia and industry.

From the early 1990s to the early 2000s, there have been extensive study on content-based image search. The progress in those years has been comprehensively discussed in existing survey papers [5] [6] [7]. Around the early 2000s, the introduction of some new insights and methods triggers another research trend in CBIR. Specially, two pioneering works have paved the way to the significant advance in content-based visual retrieval on large-scale multimedia database. The first one is the introduction of invariant local visual feature SIFT [8]. SIFT is demonstrated with excellent descriptive and discriminative power to capture visual content in a variety of literature. It can well capture the invariance to rotation and scaling transformation and is robust to illumination change. The second work is the introduction of the Bag-of-Visual-Words (BoW) model [9]. Leveraged from information retrieval, the BoW model makes a compact representation of images based on the quantization of the contained local features and is readily adapted to the classic inverted file indexing structure for scalable image retrieval.

Based on the above pioneering works, the last decade has witnessed the emergence of numerous work on multimedia content-based image retrieval [10] [11] [12] [13] [9] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29]. Meanwhile, in industry, some commercial engines on content-based image search have been launched with different focuses, such as Tineye[1], Ditto[2], Snap Fashion[3], ViSenze[4], Cortica[5], *etc*. Tineye is launched as a billion-scale reverse image search engine in May, 2008. Until January of 2017, the indexed image database size in Tineye has reached up to 17 billion. Different from Tineye, Ditto is specially focused on brand images in the wild. It provides an access to uncover the

- *Wengang Zhou and Houqiang Li are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China. E-mail: {zhwg, lihq}@ustc.edu.cn.*
- *Qi Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, 78249, USA. E-mail: qitian@cs.utsa.edu.*

1. http://tineye.com/
2. http://ditto.us.com/
3. https://www.snapfashion.co.uk/
4. https://www.visenze.com
5. http://www.cortica.com/

brands inside the shared photos on the public social media web sites.

Technically speaking, there are three key issues in content-based image retrieval: image representation, image organization, and image similarity measurement. Existing algorithms can also be categorized based on their contributions to those three key items.

Image representation originates from the fact that the intrinsic problem in content-based visual retrieval is image comparison. For convenience of comparison, an image is transformed to some kind of feature space. The motivation is to achieve an implicit alignment so as to eliminate the impact of background and potential transformations or changes while keeping the intrinsic visual content distinguishable. In fact, how to represent an image is a fundamental problem in computer vision for image understanding. There is a saying that "An image is worth a thousand words". However, it is nontrivial to identify those "words". Usually, images are represented as one or multiple visual features. The representation is expected to be descriptive and discriminative so as to distinguish similar and dissimilar images. More importantly, it is also expected to be invariant to various transformations, such as translation, rotation, resizing, illumination change, *etc*.

In multimedia retrieval, the visual database is usually very large. It is a nontrivial issue to organize the large scale database to efficiently identify the relevant results of a given query. Inspired by the success of information retrieval, many existing content-based visual retrieval algorithms and systems leverage the classic inverted file structure to index large scale visual database for scalable retrieval. Meanwhile, some hashing based techniques are also proposed for indexing in a similar perspective. To achieve this goal, visual codebook learning and feature quantization on high-dimensional visual features are involved, with spatial context embedded to further enrich the discriminative capability of the visual representation.

Ideally, the similarity between images should reflect the relevance in semantics, which, however, is difficult due to the intrinsic "semantic gap" problem. Conventionally, the image similarity in content-based retrieval is formulated based on the visual feature matching results with some weighing schemes. Alternatively, the image similarity formulations in existing algorithms can also be viewed as different match kernels [30].

In this paper, we focus on the overview over research works in the past decade after 2003. For discussion before and around 2003, we refer readers to previous survey [5] [6][7]. Recently, there have been some surveys related to CBIR [31] [2] [3]. In [31], Zhang *et al.* surveyed image search in the past 20 years from the perspective of database scaling from thousands to billions. In [3], Li *et al.* made a review of the state-of-the-art CBIR techniques in the context of social image tagging, with focus on three closed linked problems, including image tag assignment, refinement, and tag-based image retrieval. Another recent related survey is referred in [2]. In this work, we approach the recent advance in CBIR with different insights and emphasize more on the progress in methodology of a generic framework.

In the following sections, we first briefly review the generic pipeline of content-based image search. Then, we discuss five key modules of the pipeline, respectively. After that, we introduce the ground-truth datasets popularly exploited and the evaluation metrics. Finally, we discuss future potential directions and conclude this survey.

2 General Flowchart Overview

Content-based image search or retrieval has been a core problem in the multimedia field for over two decades. The general flowchart is illustrated in Fig. 1. Such a visual search framework consists of an off-line stage and an on-line stage. In the off-line stage, the database is built by image crawling and each database image is represented into some vectors and then indexed. In the on-line stage, several modules are involved, including user intention analysis, query formation, image representation, image scoring, search reranking, and retrieval browsing. The image representation module is shared in both the off-line and on-line stages. ~~This paper will not cover image crawling, user intention analysis [32], and retrieval browsing [33], of which the survey can be referred in previous work [6] [34].~~ In the following, we will focus on the other five modules, *i.e.*, query formation, image representation, database indexing, image scoring, and search reranking.

In the following sections, we make a review of related work in each module, discuss and evaluate a variety of strategies to address the key issues in the corresponding modules.

3 Query Formation

At the beginning of image retrieval, a user expresses his or her imaginary intention into some concrete visual query. The quality of the query has a significant impact on the retrieval results. A good and specific query may sufficiently reduce the retrieval difficulty and lead to satisfactory retrieval results. Generally, there are several kinds of query formation, such as query by example image, query by sketch map, query by color map, query by context map, *etc*. As illustrated in Fig. 2, different query schemes lead to significantly distinguishing results. In the following, we will discuss each of those representative query formations.

The most intuitive query formation is query by example image. That is, a user has an example image at hand and would like to retrieve more or better images about the same or similar semantics. For instance, a picture holder may want to check whether his picture is used in some web pages without his permission; a cybercop may want to check a terrorist logo appearing in the Web images or videos for anti-terrorism. To eliminate the effect of the background, a bounding box may be specified in the example image to constrain the region of interest for query. Since the example images are objective without little human involvement, it is convenient to make quantitative analysis based on it so as to guide the design of the corresponding algorithms. Therefore, query by example is the most widely explored query formation style in the research on content-based image retrieval [9] [10] [35] [36].

Besides query by example, a user may also express his intention with a sketch map [37] [38]. In this way, the query is a contour image. Since sketch is more close to the semantic

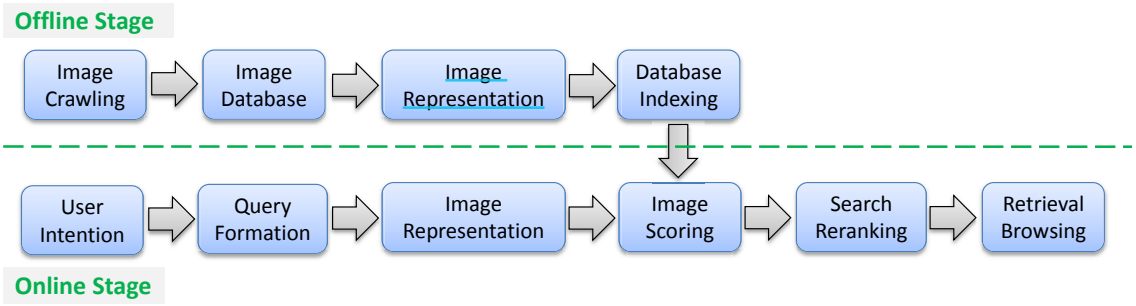

Fig. 1. The general framework of content-based image retrieval. The modules above and below the green dashed line are in the off-line stage and on-line stage, respectively. In this paper, we focus the discussion on five components, *i.e.*, query formation, image representation, database indexing, image scoring, and search reranking.

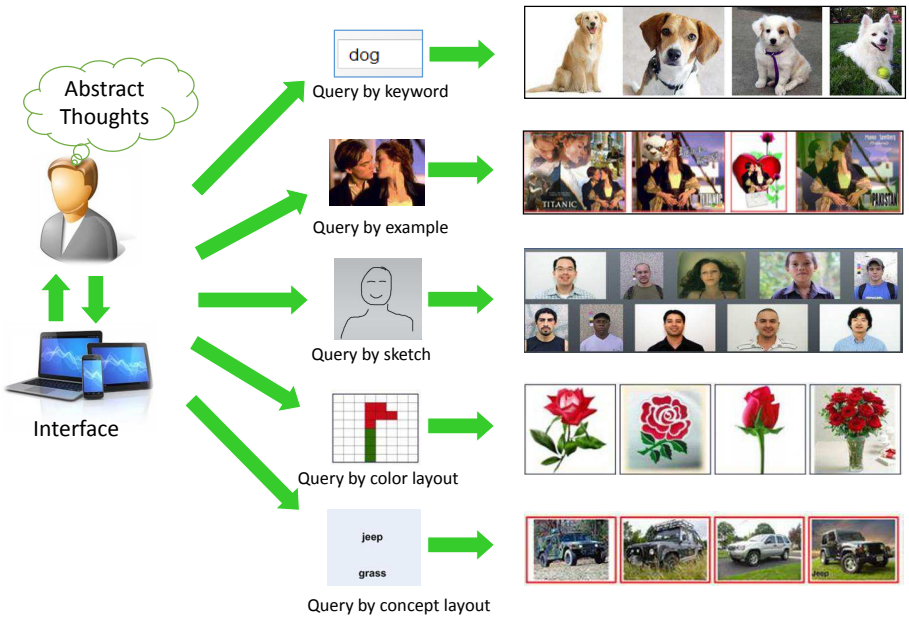

Fig. 2. Illustration of different query schemes with the corresponding retrieval results.

representation, it tends to help retrieve target results in users' mind from the semantic perspective [37]. Initial works on sketch based retrieval are limited to search for special artworks, such as clip arts [39] [40] and simple patterns [41]. As a milestone, the representative work on sketch-based retrieval for natural images is the edgel [42]. Sketch has also been employed in some image search engines, such as Gazopa[6] and Retrievr[7]. However, there are two non-trivial issues on sketch based query. Firstly, although some simple concepts, such as sun, fish, and flower, can be easily interpreted as simple shapes, in most time, it is difficult for a user to quickly sketch out what he wants to search. Secondly, since the images in the database are usually natural images, it needs to design special algorithms to convert them to sketch maps consistent with user intention.

Another query formation is color map. A user is allowed to specify the spatial distribution of colors in a given grid-like palette to generate a color map, which is used as query to retrieve images with similar colors in the relative regions of the image plain [43]. With coarse shape embedded, the color map based query can easily involve user interaction to improve the retrieval results but is limited by potential concepts to be represented. Besides, color or illumination change is prevalent in image capturing, which casts severe challenge on the reliance of color-based feature.

The above query formations are convenient for uses to input but may still be difficult to express the user's semantic intention. To alleviate this problem, Xu *et al.* proposed to form the query with concepts by text words in some specific layout in the image plain [44] [45]. Such structured object query is also explored in [46] with a latent ranking SVM model. This kind of query is specially suitable for searching generalized objects or scenes with context when the object recognition results are ready for the database images and the queries.

It is notable that, in the above query schemes taken by most existing work, the query takes the form of single image, which may be insufficient to reflect user intension in some situations. If provided with multiple probe images as query, some new strategies are expected to collaboratively represent the the query or fuse the retrieval results of each single probe [47]. That may be an interesting research topic

6. http://www.gazopa.com/
7. http://labs.systemone.at/retrievr

especially in the case of video retrieval where the query a video shot of temporal sequence.

4 Image Representation

In content based image retrieval, the key problem is how to efficiently measure the similarity between images. Since the visual objects or scenes may undergo various changes or transformations, it is infeasible to directly compare images at pixel level. Usually, visual features are extracted from images and subsequently transformed into a fix-sized vector for image representation. Considering the contradiction between large scale image database and the requirement for efficient query response, it is necessary to "pack" the visual features to facilitate the following indexing and image comparison. To achieve this goal, quantization with visual codebook training are used as a routine encoding processing for feature aggregation/pooling. Besides, as an important characteristic for visual data, spatial context is demonstrated vital to improve the distinctiveness of visual representation.

Based on the above discussion, we can mathematically formulate the content similarity between two images $\mathcal{X}$ and $\mathcal{Y}$ in Eq. 1.

$$S(\mathcal{X},\mathcal{Y}) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} k(x,y) \quad (1)$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \phi(x)^T\phi(y) \quad (2)$$

$$= \Psi(\mathcal{X})^T\Psi(\mathcal{Y}). \quad (3)$$

Based on Eq. 1, there emerge three questions.

1) Firstly, how to describe the content image $\mathcal{X}$ by a set of visual features $\{x_1, x_2, \cdots\}$?
2) Secondly, how to transform feature sets $\mathcal{X} = \{x_1, x_2, \cdots\}$ with various sizes to a fixed-length vector $\Psi(\mathcal{X})$?
3) Thirdly, how to efficiently compute the similarity between the fixed-length vectors $\Psi(\mathcal{X})^T\Psi(\mathcal{Y})$?

The above three questions essentially correspond to the feature extraction, feature encoding & aggregation, and database indexing, respectively. As for feature encoding and aggregation, it involves visual codebook learning, spatial context embedding, and quantization. In this section, we discuss the related works on those key issues in image representation, including feature extraction, visual codebook learning, spatial context embedding, quantization, and feature aggregation. The database indexing is left to the next section for discussion.

4.1 Feature Extraction

Traditionally, visual features are heuristically designed and can be categorized into local features and global features. Besides those hand-crafted features, recent years have witnessed the development of learning-based features. In the following, we will discuss those two kinds of features, respectively.

4.1.1 Hand Crafted Feature

In early CBIR algorithms and systems, global features are commonly used to describe image content by color [48] [43], shape [42] [49] [50] [51], texture [52][53], and structure [54] into a single holistic representation. As one of the representative global feature, GIST feature [55] is biologically plausible with low computational complexity and has been widely applied to evaluate approximate nearest neighbor search algorithms [56] [57] [58] [59]. With compact representation and efficient implementation, global visual feature are very suitable for duplicate detection in large-scale image database [54], but may not work well when the target images involve background clutter. Typically, global features can be used as a complementary part to improve the accuracy on near-duplicate image search based on local features [24].

Since the introduction of SIFT feature by Lowe [60] [8], local feature has been extensively explored as a routine image representation in many works on content-based image retrieval. Generally, local feature extraction involves two key steps, *i.e.* interest point detection and local region description. In interest point detection, some key points or regions with characteristic scale are detected with high repeatability. The repeatability here means that the interest points can be identified under various transformations or changes. Popular detectors include Difference of Gaussian (DoG) [8], MSER [61], Hessian affine detector [62], Harris-Hessian detector [63], and FAST [64]. In interest point detection, the invariance to translation and resizing is achieved. Distinguished from the above methods, it is also possible to obtain the interest points by uniformly and densely sample the image plane without any explicit detector [65].

After the detection of interest points, a descriptor or multiple descriptors [66] are extracted to describe the visual appearance of the local region centered at the interest point. Usually, the descriptor is designed to be invariant to rotation change and robust to affine distortion, addition of noise, and illumination changes, *etc.* Besides, it should also be distinctive so as to correctly match a single feature with high probability against a large corpus of features from many images. Such property is especially emphasized in the scenario of large-scale visual applications. The most popular choice with the above merits is SIFT feature [8]. As a variant, SURF [67] is demonstrated with comparable performance but better efficiency.

Some improvements or extensions have been made on the basis of SIFT. In [23], Arandjelovic *et al* proposed a root-SIFT by making root-normalization on the original SIFT descriptor. Although such operation is simple, it is demonstrated to significantly improve the image retrieval accuracy and can be readily plugged into many SIFT based image retrieval algorithms [68]. Zhou *et al.* proposed to generate binary signature of the SIFT descriptor with two median thresholds determined by the original descriptor itself [36]. The obtained binary SIFT leads to a new indexing scheme for image retrieval [69]. Liu *et al.* extend the binary SIFT by first generating a binary comparison matrix via dimension-pair comparison and then flexibly dividing the matrix entries into segments each of which is hashed to a bit [70]. In [21], the SIFT descriptor is transformed to

binary code with principal component analysis (PCA) and simple thresholding operations simply based on coefficients' sign. In [71], Affine-SIFT (ASIFT) simulates a set of sample views of the initial images by varying the two camera axis orientation parameters, *i.e.*, the latitude and the longitude angles and covers effectively all six parameters of the affine transformation, consequently achieving fully affine invariance.

SIFT features extracted in regions with weak internal structure suffers poor distinctiveness and may degrade image retrieval performance. To identify and remove those features, Dong *et al.* regarded a SIFT descriptor as 128 samples of a discrete random variable ranging from 0 to 255 and make use of the entropy as a measurement metric to filter SIFT features with low entropy [72].

Apart from floating point feature like SIFT, binary features are popularly explored and directly extracted from the local region of interest. Recently, binary feature BRIEF [73] and its variants, such as ORB [74], FREAK [75], and BRISK [76], have been proposed and have attracted a great deal of attention in visual matching applications. Those binary features are computed by some simple intensity difference tests, which are extremely computationally efficient. With the advantage in efficiency from Hamming distance computation, those binary features based on FAST detector [64] may have potential in large scale image search. In [77], Zhang *et al.* proposed a novel ultra short binary descriptor (USB) from the local regions of regions detected by DoG detector. The USB achieves fast image matching and indexing. Besides, following the binary SIFT scheme [36], it avoids the expensive codebook training and feature quantization in BoW model for image retrieval. A comprehensive evaluation of binary descriptors are referred in [78].

Besides the gradient information in the local regions as in SIFT feature, edge and color can also be expressed into a compact descriptor, generating Edge-SIFT [79] and color-SIFT [80]. As a binary local feature, Edge-SIFT [79] describes a local region with the extracted Canny edge detection results. Zheng *et al* extracted color name feature from the local regions, which is further transformed to a binary signature to enhance the discrimination of local SIFT feature [68].

4.1.2 Learning-based Feature

Apart from the above handcrafted visual features, it is also possible to learn features in a data-driven manner for image retrieval. Attribute feature, originally used for object categorization, can be used to represent the semantic characteristics for image retrieval [81] [82] [83]. Generally, the attribute vocabulary can be manually defined by humans [84] [85] or some ontology [86]. For each attribute, a classifier can be trained with kernel over multiple low-level visual features based on labeled training image set and used to predict the attribute score for unseen images [86] [85] [87] [88]. In [89], the attribute feature is adopted as a semantic-aware representation to compensate local SIFT feature for image search. Karayev *et al.* learned classifiers to predict image styles and applied it to search and rank image collection by styles [90]. The advantage of attribute feature is that it provides an elegant way to approximate the visual semantics so as to reduce the semantic gap. However, there are two issues on attribute features. Firstly, it is difficult to define a complete set of attribute vocabulary, either manually or in an automatic manner. Thus, the representation with the limited attribute vocabulary may be biased for a large and semantically diverse image database. Secondly, it is usually computationally expensive to extract attribute features due to the necessity to do classification over thousands of attribute categories [81] [86].

Topic models, such as probabilistic Latent Semantic Analysis (pLSA) model [91] and Latent Dirichlet Allocation (LDA) model [92], are popularly adopted to learn feature representation with semantics embedded for image retrieval [93] [94].

With the explosive research on deep neural network (DNN) [65] [95] [96], recent years have witnessed the success of the learning-based features in multiple areas. With the deep architectures, high-level abstractions close to human cognition process can be learned [97]. As a result, it is feasible to adopt DNN to extract semantic-aware features by the activations of different lays in the networks. In [98], features are extracted in local patches with a deep restricted Boltzmann machine (DBN) which is refined by using back-propagation. As a typical structure of the DNN family, deep convolutional neural network (CNN) [99] has demonstrated state-of-the-art performance in various tasks on image recognition and retrieval [100]. In [101], comprehensive studies is conducted on the potential of learned visual features with deep CNN for various applications including content based image retrieval. Razavian *et al.* study the Alex-Net [99] and VGG-Net [95], and exploit the last convolutional layers response with max pooling as image representation for image retrieval [102]. In [103], the activations of the sixth layer of the Alex-Net [99] is taken out as a DNN feature for each image, which is fused in the image similarity score level with traditional visual features including SIFT-based BoW feature, HSV histogram, and GIST.

Besides working as a global description of images, learning-based feature can also be obtained in a manner similar to local features [104]. The local regions of interest are generated by unsupervised object detection algorithms, such as selective search [105], objectness [106], and binarized normed gradients (BING) [107]. Those algorithms generate a number of object proposals in the form of bounding boxes. Then, in each object proposal region, the learning-based feature can be extracted. In [108], Sun *et al.* adopted the CNN model to extract features from local image regions detected by a general object detector [107], and applied it for image retrieval and demonstrated impressive performance. Considering the fact that object detection is sensitive to rotation transformation, Xie *et al.* proposed to rotate the test image by four different angles and then conduct object detection. Object proposals with top detection scores are then selected to extract the deep CNN feature [99]. Tolias *et al.* generate feature vector of regional maximum activation of convolutions (R-MAC) towards geometry-aware re-ranking [109]. To speedup the max-pooling operation, a novel approximation is proposed by extending the idea of integral images. In [110], the R-MAC descriptor is extended by selecting regions with a region-of-interest (ROI) selector based on region proposal network [111].

In the above approaches, the learning-based feature is extracted with the deep learning model trained for classification task. As a result, the learned feature may not well reflect the visual content characteristics of retrieval images, which may result in limited retrieval performance. Therefore, it is preferred to train the deep learning model directly for the retrieval task, which, however, is difficult since the potential image category in retrieval is difficult to define or enumerated. To partially address this difficulty, Babenko *et al.* focus on landmark retrieval and fine-tune the pre-trained CNN model on ImageNet with the class corresponding to landmarks [112]. after the fine-tuning, promising performance improvement is witnessed on the retrieval datasets with similar visual statistics, such as the Oxford Building dataset [11]. To get rid of the dependence on examples or class labels, Paulin *et al.* proposed to generate patch-level feature representation based on convolutional kernel networks in an unsupervised way [113]. In [114], the supervision takes the form of binary codes, which are obtained by decomposing the similarity matrix of training images. The resultant deep CNN model is therefore ready to generate binary codes for images in an end-to-end way. Further, Lai *et al.* propose deep neuron networks to hash images into short binary codes with optimization based on triplet ranking loss [115]. The resulted short binary codes for image representation enable efficient retrieval by Hamming distance and considerable gain in storage.

4.2 Visual Codebook Learning

Usually, hundreds or thousands of local features can be extracted from a single image. To achieve a compact representation, high dimensional local features are quantized to visual words of a pre-trained visual codebook, and based on the quantization results an image with a set of local features can be transformed to a fixed-length vector, by the Bag-of-Visual-Words model [9], VLAD [116], or Fisher Vector [117]. To generate a visual codebook beforehand, the most intuitive way is by clustering the training feature samples with brute-force k-means [9] [12] and then regarding the clustering centers as visual words. Since the local feature dimension is high and the training sample corpus is large, it suffers extremely high computational complexity to train a large, say, million-scale or larger, visual codebook. To address this problem, an alternative to to adopt the hierarchical k-means [10], which reduces the computational complexity from linear to logarithm for large size visual codebook generation.

In the standard k-means, the most computing overhead is consumed on the assignment of feature samples to the close cluster center vector, which is implemented by linearly comparing all cluster center vectors. That process can be speeded up by replacing the linear scan with approximate nearest neighbor search. With such observation, Philbin *et al.* proposed an approximate k-means algorithm by exploiting randomized k-D trees for fast assignment [11]. Instead of using k-means to generate visual words, Li *et al.* generated hyper-spheres by randomly sampling seed feature points with a predefined radius [118]. Then, those hyper-spheres with the seed features corresponds to the visual codebook. In [119], Chu *et al.* proposed to build the visual vocabulary based on graph density. It measures the intra-word similarity by the feature graph density and derives the visual word by dense feature graph with a Scalable Maximization Estimation (SME) scheme.

In the Bag-of-Visual-Words model, the visual codebook works as a media to identify the visual word ID, which can be regarded as the quantization or hashing result. In other words, it is feasible to directly transform the visual feature to a visual word ID without explicitly defining the visual word. Following this idea, different from the above codebook generation methods, some other approaches on image retrieval generate a virtual visual codebook without explicit training. Those methods transform a local feature to binary signature, based on which the visual word ID is heuristically defined. In [21], Zhang *et al.* proposed a new query-sensitive ranking algorithm to rank PCA-based binary hash codes to search for ϵ-neighbors for image retrieval. The binary signature is generated with a LSH (locality sensitive hashing) strategy and the top bits are used as the visual word ID to group feature points with the same ID. Zhou *et al.* [36] proposed to binarize a SIFT descriptor into a 256-bit binary signature. Without training a codebook, this method selects 32 bits from the 256-bit vector as a codeword for indexing and search. The drawback of this approach is that the rest 224 bits per feature have to be stored in the inverted index lists, which casts a heavy overhead on memory. Similarly, Dong *et al* proposed to transform to a SIFT descriptor to a 128-bit vector [72] with a sketch embedding technique [120]. Then, the 128-bit vector is divided into 4 non-overlapped block, each of which is considered as a key or a visual word for later indexing. In [121], Zhou *et al* proposed a codebook-training-free framework based on scalable cascaded hashing. To ensure the recall rate of feature matching, the scalable cascaded hashing (SCH) scheme which conducts scalar quantization on the principal components of local descriptors in a cascaded manner.

4.3 Spatial Context Embedding

As the representation of structured visual content, visual features are correlated by spatial context in terms of orientation, scale, and key points' distance in image plane. By including the contextual information, the discriminative capability of visual codebook can be greatly enhanced [26]. Analogy to the text phrase in information retrieval, it is feasible to generate visual phrase over visual words. In [27] [122], neighboring local features are correlated to generate high-order visual phrases, which are further refined to be more descriptive for content representation.

Many algorithms target on modeling the local spatial context among local visual features. Loose spatial consistency from some spatially nearest neighbors can be imposed to filter false visual-word matches. Supports are collected by checking the matched features with the search area defined by 15 nearest neighbors [9]. Such loose scheme, although efficient, is sensitive to the image noise incurred by editing. Zhang *et al.* generated contextual visual codebook by modeling the spatial context of local features in group with a discriminant group distance metric [28]. Wang *et al.* proposed descriptor contextual weighting (DCW) and spatial contextual weighting (SCW) of local features in the descriptor domain and spatial domain, respectively, to upgrade

the vocabulary tree based approach [123]. DCW down-weights less informative features based on frequencies of descriptor quantization paths on a vocabulary tree while SCW exploits some efficient spatial contextual statistics to preserve the rich descriptive information of local features. In [124], Liu *et al.* built a spatial-relationship dictionary by embedding spatial context among local features for image retrieval.

Further, the multi-modal property that multiple different features are extracted at an identical key points is discussed and explored for contextual hashing [125]. In [126], geometric min-hashing constructs repeatable hash keys with loosely local geometric information for more discriminative description. In [17], Wu *et al.* proposed to bundle local features in a MSER region [61]. The MSER regions are defined by an extremal property of the intensity function in the region and on its outer boundary and are detected as stable regions across a threshold range from watershed-based segmentation [61]. Bundled features are compared by the shared visual word amount and the relative ordering of matched visual words. In [63], ordinal measure (OM) feature [127] is extracted from the spatial neighborhood around local interest points. Then, local spatial consistency verification is conducted by checking whether the OMs of the correspondence features are below a predefined threshold.

Different from the above approaches, Cao *et al.* modeled the global spatial context by two families of ordered bag-of-features as a generation of the spatial pyramid matching [128] by linear projection and circular projection and further refined them to capture the invariance of object translation, rotation, and scaling by simple histogram operations, including calibration, equalization, and decomposition [129].

In the scenario of face retrieval, the above general codebook generation methods are likely to fail to capture the unique facial characteristics. To generate discriminative visual codebook, Wu *et al.* proposed to generate identity-based visual vocabulary with some training persons each with multiple face examples under various poses, expressions, and illumination conditions [130]. A visual word is defined as a tuple consisting of two components, *i.e.*, person ID and position ID and associated with multiple examples.

4.4 Feature Quantization

With visual codebook defined, feature quantization is to assign a visual word ID to each feature. To design a suitable assignment function, special consideration should be made to balance quantization accuracy, efficiency, and memory overhead.

The most naive choice is to take the nearest neighbor search, so as to find the closest (the most similar) visual word of a given feature by linear scan, which, however, suffers expensive computational cost. Usually, approximate nearest neighbor (ANN) search methods are adopted to speed up the searching process, with sacrifice of accuracy to some extent. In [8], a k-d tree structure [131] is utilized with a best-bin-first modification to find approximate nearest neighbors to the descriptor vector of the query. In [10], based on the hierarchical vocabulary tree, an efficient approximate nearest neighbor search is achieved by propagating the query feature vector from the root node down the tree by comparing the corresponding child nodes and choosing the closest one. In [132], a k-d forest approximation algorithm is proposed with reduced time complexity. Muja and Lowe proposed a novel priority search k-means tree algorithm for scalable nearest neighbor search [133] with FLANN library[8] provided. In [118], the feature quantization is achieved by range-based neighbor search over the random seeding codebook. This random seeding approach, although efficient in implementation, suffers the bias to the training data and achieves limited retrieval accuracy in large-scale image retrieval [134]. Those approaches conduct quantization in a hard manner and inevitably incur severe quantization loss.

Considering that the codebook partitions the feature space into some non-overlapping cells, feature quantization works to identify which cell a test feature falls into. When the codebook size is large which means the feature space is finely partitioned, features proximate to the partition boundary are likely to fall into different cells. On the other hand, with small codebook and feature space coarsely partitioned, irrelevant features with large distance may also fall into the same cell. Both cases will incur quantization loss and degrade the recall and precision of feature matching, respectively. A trade-off shall be made on the codebook size to balance the recall and precision from the above two kinds of loss [10], or some constraints are involved to refine the quantization quality.

Some approaches adopt a large visual codebook but take account of the soft quantization to reduce the quantization loss. Generally, a descriptor-dependent soft assignment scheme [15] is used to map a feature vector to a weighted combination of multiple visual words. Intuitively, the soft quantization can be performed for both a query feature and the database features. However, it will cost several times more memory to store the multiple quantization results for each database feature. As a trade-off, the soft quantization can be constrained to only the query side [35]. In [35], a new quantizer is designed based on a codebook learned by brute-force k-means clustering. It first performs k-means clustering on the pre-trained visual words and generate a two-layer visual vocabulary tree in a bottom-up way. Then, new connections between the two-layer nodes are constructed by quantizing a large feature set with both layers of quantizers. Soft assignment is performed with a criteria based on distance ratio.

On the other hand, some other approaches keep a relatively small visual codebook but performs further verification to reduce the quantization loss. In [12], Hamming Embedding reduces the dimension of SIFT descriptors quantized to a visual word, and then trains a median vector by taking the median value in each dimension of the feature samples. After a new feature is quantized to a visual word, it is projected to the low dimensional space, and then compared with the median vector dimension-wise to generate binary signature for matching verification [54]. In [135], a variant, *i.e.*, the asymmetric Hamming Embedding scheme, is proposed to exploit the rich information conveyed by the binary signature. Zhou *et al.*adopt a similar verification

8. http://www.cs.ubc.ca/research/flann/

idea with a different binary signature which is generated by comparing each element of a feature descriptor to its median [136].

The above approaches rely on single visual codebook for feature quantization. To correct quantization artifacts and improve recall, typically, multiple vocabularies are generated for feature quantization to improve the recall [137][138]. Since multiple vocabularies suffers from vocabulary correlation, Zheng *et al* proposed a Bayes merging approach to down-weight the indexed features in the intersection set of multiple vocabularies [139]. It models the the correlation problem in a probabilistic view and estimate a joint similarity on both image- and feature-level for the indexed features in the intersection set.

The vector quantization of local descriptors is closely related to approximate nearest neighbor search [58]. In literature, there are many hashing algorithms for approximate nearest neighbor (ANN) search, such as LSH [140][141], multi-probe LSH [142], kernelized LSH [56], semi-supervised hashing method (SSH) [143], spectral hashing [57], min-Hashing [16], iterative quantization [144], random grids [145], bucket distance hashing (BDH) [146], query-driven iterated neighborhood graph search [147], and linear distance preserving hashing [148]. These hashing methods, however, are mostly applied to global image features such as GIST or BoW features at the image level, or to feature retrieval only at the local feature level. There is few work dedicated to image level search based on local feature hashing [22]. The major concern of those hashing methods is that multiple hashing tables are usually involved and each feature needs to be indexed multiple times, which cast heavy memory burden. Besides, in hashing methods such as LSH [141], multi-probe LSH [142] and kernelized LSH [56], the original database feature vectors need be kept in memory to compute the exact distance to the query feature, which is infeasible in the scenario of large-scale image search with local features. Moreover, approximate nearest neighbor search usually targets at identifying the top-k closest data to the query, which ignores the essence of range-based neighbor search in visual feature matching. That is, given a query feature, the number of target data in the database is query-sensitive and determined by the coverage of the range-based neighborhood of the query.

In [58], a novel product quantization is proposed to generate an exponentially large codebook with low cost in memory and time for approximate nearest neighbor search. It decomposes the feature space into a Cartesian product of low-dimensional subspaces and quantizes each sub-space individually. The quantization indices of each sub-space are presented as a short code, based on which the Euclidean distance between two feature vectors can be efficiently estimated by looking up a pre-computed table. The product quantization, however, suffers from exhaustive search for identifying target features, which is prohibitive in large-scale image search [58]. As a partial solution to this bottle neck, vector quantization by k-means can be involved to narrow the search scope and allow the product to focus on a small fraction of indexed features [58]. In [149], the product quantization is optimized with respect to the vector space decomposition and the quantization codebook with two solutions from the non-parametric and the parametric perspectives. Zhou *et al.* formulated the feature matching as an $\epsilon-$neighborhood problem and approximated it with a dual-resolution quantization scheme for efficient indexing and querying [134]. It performs scalar quantization in coarse and fine resolutions on each dimension of the data, and cascades the quantization results over all dimensions. The cascaded quantization results in coarse resolution are used to build the index, while the cascaded quantization results in fine resolutions are transformed to a binary signature for matching verification.

In [150], the high dimensional SIFT descriptor space is partitioned into regular lattices. Although demonstrated to work well in image classification, in [15], regular lattice quantization is revealed to work much worse than [10] [15] in large scale image search application.

4.5 Feature Aggregation

When an image is represented by a set of local features, it is necessary to aggregate those local features into a fixed-length vector representation for convenience of similarity comparison between query and database images. Generally, there are three alternatives to achieve this goal. The first one is the classic Bag-of-Visual-Words representation, which quantizes each local feature to the closest visual word of a pre-trained visual codebook. The quantization result of a single local feature can be regarded as a high-dimensional binary vector, where the non-zero dimension corresponds to the quantized visual word. By pooling the quantization results of all local features in an image, we obtain a BoW vector with the dimension size as the visual codebook size. In this scheme, the involved visual codebook is usually very large in size and the generated BoW vector is very sparse, which facilitates the use of the inverted file indexing.

The second popular feature aggregation method is the VLAD (vector of locally aggregated descriptors) approach [116], which adopts k-means based vector quantization and accumulates the quantization residues for features quantized to each visual word and concatenate those accumulated vectors into a single vector representation. With compact size, the VLAD vector inherits some important properties from SIFT feature, including invariance to translation, rotation, and scaling. In [151], the VLAD approach is improved by a new intra-normalization scheme and multiple spatial VLAD representation. An in-depth analysis on VLAD is conducted in [152]. In [153], an extension of VLAD is proposed with triangulation embedding scheme and democratic aggregation technique. Further, Tolias *et al.* encompassed the VLAD vector with various matching schemes [30]. To reduce the computational complexity of the democratic aggregation scheme, Gao *et al.* proposed a fast scheme with comparable retrieval accuracy performance [154]. In [155], sparse coding is adopted to encode the local feature descriptors into sparse vectors, which are further aggregated with a max-pooling strategy. Liu *et al.* proposed a hierarchical scheme to build the VLAD vector with SIFT feature [156]. By involving a hidden-layer vocabulary, the distribution of the residue vectors to be aggregated becomes much more uniform, leading to better discrimination for the representation.

Although compact and efficient representation is achieved by global aggregation of all local features in an

image, the original VLAD vector sacrifices the flexibility to address partial occlusion and background clutter. To alleviate this problem, Liu *et al.* [157] grouped local key points by their spatial positions in the image plane and aggregated all local descriptors in each group by the VLAD scheme [116]. As a result, a local aggregation of local features is achieved and promising retrieval accuracy is demonstrated with a tradeoff in memory cost.

Besides the BoW representation and the VLAD, another alternative is the Fisher Vector based representation [117] with Fisher kernel [158] [159]. As a generative model, given a set of features for an image, Fisher vector represents them into a fix-sized vector by the gradient of the log-likelihood function with respect to a set of parameter vectors [160]. In [117] [161], Gaussian Mixture Model (GMM) is adopted as a generative model to aggregate the normalized concatenated gradient vectors of all local descriptors into a uniform Fisher vector with an average pooling scheme. In fact, the Fisher Vector can be regarded as a generalized representation of the BoW representation and VLAD. On one hand, if we keep only the gradient of the log-likelihood function with respect to the weight of GMM, the Fisher Vector degenerates to a soft version of the BoW vector. On the other hand, If we keep only the gradient of the log-likelyhood function with respect to the mean vector of GMM, we can derive the VLAD representation [58].

In either the Fish vector or VLAD representation, the involved GMM number or codebook size is relative small and the obtained aggregated vector is no long sparse. As a result, it is unsuitable to apply the inverted file indexing scheme to index images based on the aggregated results. To address this dilemma, the aggregated vector is dimensionally reduced and further encoded by product quantization [58] for efficient distance computation.

The above aggregation schemes are based on local hand-crafted feature, such as SIFT feature. Intuitively, such schemes can be directly leveraged to local deep features. Following this idea, Gong *et al.* [162] extract local CNN features from the local patches sampled regularly at multiple scale levels and pool the CNN features in each scale level with the VLAD scheme [37]. In [163], Babenko *et al.* interpret the activations from the last convolutional layers of CNNs as local deep features. They reveal that the individual similarity of local deep feature is very discriminative and the simple aggregation with sum pooling over local deep feature yields the best performance.

5 Database Indexing

Image index refers to a database organizing structure to assist for efficient retrieval of the target images. Since the response time is a key issue in retrieval, the significance of database indexing is becoming increasingly evident as the scale of image database on the Web explosively grows. Generally, in CBIR, two kinds of indexing techniques are popularly adopted, *i.e.*, inverted file indexing and hashing based indexing. In the following, we will discuss related retrieval algorithms in each category, respectively.

5.1 Inverted File Indexing

Inspired by the success of text search engines, inverted file indexing [164] has been successfully used for large scale image search [9] [11] [18] [14] [10] [12] [17] [165]. In essence, the inverted file structure is a compact column-wise representation of a sparse matrix, where the row and the column denote image and visual word, respectively. In on-line retrieval, only those images sharing common visual words with the query image need to be checked. Therefore, the number of candidate images to be compared is greatly reduced, achieving an efficient response.

In the inverted file structure, each visual word is followed by an inverted file list of entries. Each entry stores the ID of the image where the visual word appears, and some other clues for verification or similarity measurement. For instance, Hamming Embedding [12] generates a 64-bit Hamming code for each feature to verify the descriptor matching. The geometric clues, such as feature position, scale, and orientation, are also stored in the inverted file list for geometric consistency verification [11] [18] [12] [13]. In [17], Wu *et al.* recorded the feature orders in horizontal and verification direction in each bundled feature located in a MSER region. In [123], 3 spatial statistics, including descriptor density, mean relative log scale, and mean orientation difference, are calculated for each feature and stored in the inverted list after quantization. Zheng *et al.* modeled the correlation between multiple features with a multi-IDF scheme and coupled the binary signatures of those features into the inverted file to enhances the quality of visual matching [166].

Following the general idea of inverted file structure, many variants are proposed. In [42], to adapt to the inverted index structure for sketch-based retrieval, it regularly quantizes the edge pixel in position channel and orientation channel and follows each entry in the edgel dictionary with an inverted lists of related images. In [68], Zheng *et al* proposed a new coupled Multi-Index (c-MI) framework to fuse complementary features at indexing level. Each dimension of c-MI corresponds to one kind of feature, and the retrieval process votes for images similar in both SIFT and color attribute [85] feature spaces. In [70], the image database is cross-indexed in both the binary SIFT space and the original SIFT space. With such cross-indexing structure, a new searching strategy is designed to find target data for effective feature quantization.

Some methods try to embed the semantics into the index structure. In [167], Zhang *et al* proposed a new indexing structure by decomposing a document-like representation of an image into two components, one for dimension reduction and the other for residual information preservation. The decomposition is achieved by either a graphical model or a matrix factorization approach. Then, the similarity between images is transferred to measuring similarities of their components. In [89], Zhang *et al* proposed a semantic-aware co-indexing to jointly embed two strong cues, *i.e.*, local SIFT feature and semantic attributes, into the inverted indexes. It exploits 1000 semantic attributes to filter out isolated images and insert semantically similar images to the initial inverted index set built based on local SIFT features. As a result, the discriminative capability of the indexed features is significantly enhanced.

To adapt the product quantization [58] to the inverted index idea, inverted multi-index is proposed to generalize the inverted index idea by replacing the standard quantization

within inverted indices with product quantization, so as to speed up the approximate nearest neighbor search.

To improve the recall rate of inverted indexing algorithms, the database images are indexed multiple times with multiple quantizers, such as randomized k-d trees [168] [66]. In [137], a joint inverted indexing algorithm is proposed, which jointly optimizes all codewords in all quantizers and demonstrates considerable improvement over methods with multiple independent quantizers. In [23], this goal is achieved by augmenting the image features for the database images which are estimated to be visible in a homograpy in the augmented images.

To speedup the online retrieval process, Zheng *et al.* proposed a novel Q-Index structure based on the inverted index organization [169]. It defines an impact score for each indexed local SIFT feature based on TF-IDF, scale, saliency, and quantization ambiguity. Then, based on the impact score, it introduced two complementary strategies, *i.e.* query pruning and early termination, with the former to discard less important features in the query and the later to partially visit the index lists containing the most important indexed features. The proposed algorithm demonstrates significant speed-up for online query with competitive retrieval accuracy. In [170], Ji *et al.* considered the scenario of parallelized image retrieval and proposed to distribute visual indexing structure over multiple servers. To reduce the search latency across servers, it formulates the index distribution problem as a learning problem by maximizing the uniformity of assigning the words of a given query to multiple servers.

5.2 Hashing Based Indexing

When the image representation, for instance GIST feature and VLAD feature, is a dense vector with the majority of the coefficients being non-zero, it is unsuitable to directly apply the inverted file structure for indexing. To achieve efficient retrieval for relevant results, hashing techniques are popularly adopted [171] [172] [173] [174] [175]. The most representative hashing scheme is the locality sensitive hashing (LSH) [176], which partitions the feature space with multiple hash functions of random projections with the intuition that for objects which are close to each other, the collision probability is much higher than for those which are far away. Given a query, some candidates are first retrieved based on hashing collision and re-ranked based on the exact distance from the query. In [56], LSH is generated to accommodate arbitrary kernel functions, with sub-linear time approximate similarity search permitted. The potential concern of those hashing scheme is that, since the raw database representation vectors should be stored in memory for the reranking stage, they are not well scalable to large-scale image database. In [177], a feature map is proposed by integrating appearance and global geometry, which is further hashed for indexing. This scheme, however, suffers expensive memory cost which is quadratic in the number of local features, which limits its scalability towards large scale image retrieval. To address this drawback, an extension is made with a feature selection model to replace the hashing approach [178].

With the inverted index structure, the memory cost is proportional to the amount of non-zero elements in the representation vector. To further reduce such memory overhead, Jegou *et al.* proposed to approximate the original visual word occurrence vector by projecting it onto a set of pre-defined sparse projection functions, generating multiple min-BOF descriptors [179]. Those min-BOF descriptors is further quantized for indexing. With similar attempt, in [16][180], min-Hash is proposed to describe images by mapping the visual word occurrence vector to a low-dimensional representation by a group of min-hash functions and define image similarity as the visual word set overlap. Consequently, only a small constant amount of data per image need to be stored. The potential concern of min-hashing [16][180] and its variant [126] is that although high retrieval precision can be achieved, the retrieval recall performance may be limited unless many more hashing tables are involved, which, however, imposes severe memory burden.

6 Image Scoring

In multimedia retrieval, the target results in the index image database are assigned with a relevance score for ranking and then returned to users. The relevance score can be defined either by measuring distance between the aggregated feature vectors of image representation or from the perspective of voting from relevant visual feature matches.

6.1 Distance Based Scoring

With feature aggregation, an image is represented into a fix-sized vector. The content relevance between images can be measured based on the L_p-normalized distance between their feature aggregation vectors, as shown in Eq. 4.

$$D(I_q, I_m) = \left(\sum_{i=1}^{N} |q_i - m_i|^p \right)^{\frac{1}{p}} \tag{4}$$

where the feature aggregation vectors of image I_q and I_m are denoted as $[q_1, q_2, \cdots, q_N]$ and $[m_1, m_2, \cdots, m_N]$, respectively, and N denotes the vector dimension. In [10], it is revealed that L_1-norm yields better retrieval accuracy than L_2-norm with the BoW model. Lin *et al.* extended the above feature distance to measure partial similarity between images with an optimization scheme [181].

When the BoW model is adopted for image representation, the feature aggregation vector is essentially a weighted visual word histogram obtained based on the feature quantization results. To distinguish the significance of visual words in different images, term frequency (TF) and inverted document/image frequency (IDF) are widely applied in many existing state-of-the-art algorithms [10][12][9][15][17]. Generally, the visual word vector weighted by TF and IDF are L_p-normalized for later distance computation. When the codebook size is much larger than the local feature amount in images, the aggregated feature vector of image is very sparse and we only need to check those visual words appearing in both images as illustrated in Eq. 6 [10], which is very efficient in practical implementation.

$$D(I_q, I_m) = \sum_{i=1}^{N} |q_i - m_i|^p \quad (5)$$

$$= 2 + \sum_{i|q_i \neq 0, m_i \neq 0} (|q_i - m_i|^p - q_i^p - m_i^p) \quad (6)$$

However, the dissimilarity measure by the L_p-distance is not optimal. As revealed in [182], there exists the neighborhood reversibility issue, which means that an image is usually not the k-nearest neighbor of its k-nearest neighbor images. Such issue causes that problem that some images are frequently returned while others are rarely returned when submitting query images. To address this problem, Jegou *et al.* proposed a novel contextual dissimilarity measure to refine the Euclidean distance based distance [182]. It modifies the neighborhood structure in the BoW space by iteratively estimating distance update terms in the spirit of Sinkhorns scaling algorithm. Alternatively, in [183], a probabilistic framework is proposed to model the feature to feature similarity measure and a query adaptive similarity is derived. Different from the above approaches, in [184], the similarity metric is implicitly learnt with diffusion processes by exploring the affinity graphs to capture the intrinsic manifold of database images.

In [138], Jegou *et al.* investigated the phenomenon of co-missing and co-occurrence in the regular BoW vector representation. The co-missing phenomenon denotes a negative evidence, *i.e.*, a visual word is jointly missing from two BoW vectors. To include the under-estimated evidence for similarity measurement refinement, vectors of images are centered by mean substraction [138]. On the other hand, the co-occurrence of visual words across BoW vectors will cause over-counting of some visual patterns. To limit this impact, a whitening operation is introduced to the BoW vector to generate a new representation [138]. Such preprocessing also applies to the VLAD vector [116]. Considerable accuracy gain has been demonstrated with the above operations.

6.2 Voting Based Scoring

In local feature based image retrieval, the image similarity is intrinsically determined by the feature matches between images. Therefore, it is natural to derive the image similarity score by aggregating votes from the matched features. In this way, the similarity score is not necessarily normalized, which is acceptable considering the nature of visual ranking in image retrieval.

In [13], the relevance score is simply defined by counting how many pairs of local feature are matches across two images. In [35], Jegou *et al* formulated the scoring function as a cumulation of squared TF-IDF weights on shared visual words, which is essentially a BOF (bag of features) inner product [35]. In [17], the image similarity is defined as the sum of the TF-IDF score [20], which is further enhanced with a weighting term by matching bundled feature sets. The weighting term consists of membership term and geometric term. The former term is defined as the number of shared visual words between two bundled features, while the latter is formulated using relative ordering to penalize geometric inconsistency of the matching between two bundled features. In [185][186], Zheng *et al* propose a novel Lp-norm IDF to extend the classic IDF weighting scheme.

The context clues in the descriptor space and the spatial domain are important to contribute the similarity score when comparing images. In [123], a contextual weighting scheme is introduced to enhance the original IDF-based voting so as to improve the classic vocabulary tree approach. Two kinds of weighting scheme, *i.e.*, descriptor contextual weighting (DCW) and spatial contextual weighting, are formulated to multiply the basic IDF weight as a new weighting scheme for image scoring. In [187], Shen *et al.* proposed a spatially-constrained similarity measure based on a certain transformation to formulate voting score. The transformation space is discretized and a voting map is generated based on the relative locations of matched features to determine the optimal transformation.

In [179], each indexed feature is embedded with a binary signature and the image distance is defined as a summation of the hamming distance between matched features, of which the distance for the unobserved match is set as statistical expectation of the distance. Similar scoring scheme for the unobserved match is also adopted by Liu *et al.* [157]. In [63], to tolerate the correspondences of multiple visual objects with different transformations, local similarity of deformations is derived from the peak value in the histogram of pairwise geometric consistency [188]. This similarity score is used as a weighting term to the general voting scores from local correspondences.

In image retrieval with visual word representation, similar to text-based information retrieval [189], there is a phenomenon of visual word burstiness, *i.e.*, some visual element appears much more frequently in an image than the statistically expectation, which undermines the visual similarity measure. To address this problem, Jegou *et al* proposed three strategies to penalize the voting scores from the bursting visual words by removing multiple local matches and weaken the influence of intra- and inter-images bursts [190] [191].

7 Search Reranking

The initially returned result list can be further refined by exploring the visual context [192], [193] or enhancing the original query. Geometric verification [11] [18] [12] [13] [126] [194], query expansion [14] [195], and retrieval fusion [24] are three of the most successful post-processing techniques to boost the accuracy of large scale image search. In the following, we will review the related literature in each category.

7.1 Geometric Context Verification

In image retrieval with local invariant features, the feature correspondences between query and database images are built based on the proximity of local features in the descriptor space. As a popular criteria, a tentative correspondence is built if the corresponding two local features are quantized to the same visual word of a pre-trained visual vocabulary. However, due to the ambiguity of local descriptor and the quantization loss, false correspondences of irrelevant visual

content are inevitably incurred, which confuse the similarity measurement for images and degrade the retrieval accuracy. Note that, besides the descriptor, local invariant features are characterised by other geometric context, such as the location of key points in image plane, orientation, scale, and spatial co-occurrences with other local features. Such geometric context is an important clue to depress or exclude those false matches.

Generally, among the inliers in the correspondences set, there is an underlying transformation model. If the model is uncovered, we can easily distinguish the inliers from the outliers. To model the transformation of visual object or scene across images, an affine transformation model with six parameters can be used, which estimates the rotation, scaling, translation, and perspective change in a single homography [11]. For some difficult cases, there may exist multiple homographies which makes the model estimation problem much more challenging.

Some approaches estimate the transformation model in an explicit way to verify the local correspondences. Those methods are either based the RANSAC-like idea [11][8][196] [63] or follow the Hough voting strategy [8][197]. The key idea of RANSAC [198] is to generate hypotheses on random sets of correspondences and identify a geometric model with the maximum inliers. Statistically speaking, the genuine model can be recovered with sufficient number of correspondence sampling and model evaluation. However, when the rate of inliers is small, the expected number of correspondence sampling is large, which incurs high computational complexity. In [11], by adopting the region shape of local feature, a hypothesis is generated with single correspondence, which make it feasible to enumerate all hypotheses and significantly reduces the computational cost compared with RANSAC. There are two issues on the RANSAC based algorithms. Firstly, it needs a parameter for hypothesis verification, which is usually defined in an ad-hoc way. Secondly, the computational complexity is quadratic with respect to the number of correspondences, which is somewhat expensive.

An alternative to the RANSAC-like methods, Hough voting strategy [8] [199] operates in a transformation space. In this case, the voting operation is linear to the correspondence number. In [12], the Hough voting is conducted in the space of scale and orientation. Based on the SIFT feature correspondences between images, it builds two histograms on the orientation difference and scale difference separately. Assuming that truly matched features will share similar orientation difference, it identifies the peak points in the histogram on orientation difference of matched features and regard those feature pairs with orientation difference far from the peak as irrelevant and false matches. Similar operation is also performed on the scale difference of matched features to further remove those geometrically inconsistent SIFT matches. In [20], Zhang *et al.* built a 2D Hough voting space based on the relative displacements of corresponding local features to derive the geometric-preserving visual phrase (GVP).This algorithm can be extended to address the transformation invariance to scale and rotation with the price of high memory overhead to maintain the Hough histograms. The potential problem in Hough voting is the flexibility issue in the definition of the bin size for the transformation space partition. To address the problem, in [197], motivated by the pyramid matching scheme [200], Tolias *et al.* propose a Hough pyramid matching scheme. It approximates affinity by bin size and group the correspondences based on the affinity in a bottom-up way. Notably, the complexity of this algorithm is linear to the correspondence number. In [199], the Hough pyramid matching scheme is extended by including the soft assignment for feature quantization on the query image. Different from the above methods, Li *et al.* proposed a novel pairwise geometric matching method [194] for implicit spatial verification at a significantly reduced computational cost. To reduce the correspondence redundancy, it first builds the initial correspondence set with a one-versus-one matching strategy, which is further refined based on Hough voting in the scaling and rotation transformation space [12]. Based on the reliable correspondence set, a new pairwise weighting method is proposed to measure the matching score between two images.

Some other algorithms approach the geometric context verification problem without explicit handling the transformation model. Sivic *et al.* adopted the consistency of spatial context in local feature groups to verify correspondences [9]. In [18], a spatial coding scheme is proposed to encode into two binary maps by comparing the relative coordinates of matched feature points in horizontal and vertical directions, respectively. Then it recursively removes geometrically inconsistent matches by analyzing those maps. Although spatial coding map is invariant to image changes in translation and scaling, it cannot handle the rotation change. In [13] [201], Zhou *et al.* extended the spatial coding by including the characteristic orientation and scale of SIFT feature and proposed two geometric context coding methods, *i.e.*, geometric square coding and geometric fan coding. The geometric coding algorithm can well handle image changes in translation, rotation, and scaling. In [202], Chu *et al.* proposed a Combined-Orientation-Position (COP) consistency graph model to measure the relative spatial consistency among the candidate matches of SIFT features with a coarse-to-fine family of evenly sectored polar coordinate system. Those spatially inconsistent noisy features are effectively identified and rejected by detecting the group of candidate feature matches with the largest average COP consistency.

7.2 Query Expansion

Query expansion, leveraged from text retrieval, reissues the initially highly-ranked results to generate new queries. Some relevant features, which are not present in the original query, can be used to enrich the original query to further improve the recall performance. Several expansion strategies, such as average query expansion, transitive closure expansion, recursive average query expansion, intra-expansion, and inter-expansion, *etc.*, have been discussed in [14] [195].

In [23], a discriminative query expansion algorithm is proposed. It takes spatially verified images as positive data and images with low tf-idf scores as the negative training data. Then, a classifier is learnt on-the-fly and images are sorted by their signed distances from the decision boundary. In [203], Xie *et al.* constructed a sparse graph by connecting

potentially relevant images offline and adopted a query-dependent algorithm, *i.e.*, HITS [204], to reranking images based on affinity propagation. Further, Xie *et al.* formulated the search process with a heterogeneous graph model and proposed two graph-based re-ranking algorithms to improve the search precision and recall, respectively [205]. It first incrementally identifies the most reliable images from the database to expand the query so as to boost the recall. After that, an image-feature voting scheme is used to iteratively update the scores of images and features to re-rank images. In [206], a contextual query expansion scheme is proposed to explore the common visual patterns. The contextual query expansion is performed in both the visual word level and the image level.

As a special case of query expansion, relevance feedback [1] has been demonstrated to be successful search re-ranking technique and well studied before 2000 and received some attention in recent years [207] [208] [209] [210] [211] [212]. In relevance feedback, the key idea is to learn a query-specific similarity metric based on the relevant and irrelevant examples indicated by users. Some discriminative models are learned with SVM [207][208] or boosting schemes [213]. Considering that users are usually reluctant or impatient to specify positive or negative images, user click log information can be collected as feedback to implicitly improve the retrieval system [31] [214]. For more discussion on relevance feedback, we refer readers to [215] [216] for a comprehensive survey.

7.3 Retrieval Fusion

An image can be represented by different features, based on which different methods can be designed for retrieval. If the retrieval results of different methods are complementary to each other, they can be fused to obtain better results. Most approaches conduct retrieval fusion in the rank level. Fagin *et al.* proposed a rank aggregation algorithm to combine the image ranking lists of multiple independent retrieval methods or "voters" [217]. In [24], the retrieval fusion is formulated as a graph-based ranking problem. A weighted undirected graph is built based on the retrieval results of one method and the graphs corresponding to multiple retrieval methods are fused to a single graph, based on which, link analysis [218] or maximizing weighted density is conducted to identify the relevance score and rank the retrieval results. In [219], Ye *et al.* proposed a novel rank minimization method to fuse the confidence scores of multiple different models. It first constructs a comparative relationship matrix based on the predicted confident scores for each model. With the assumption that the relative score relations are consistent across different models with some sparse deviations, it formulates the score fusion problem as seeking a shred rank-2 matrix and derives a robust a score vector.

Different from the above fusion methods, Zheng *et al.* approached the retrieval fusion in the score level [103]. Motivated by the shape differences in the ranked score curve between good and bad representation features, it normalizes the score curves by reference curves trained on irrelevant data and derives an effectiveness score based on the area under the normalized score curve. Then, the query similarity measurement is adaptively formulated in a product manner over the feature scores weighted by the effectiveness score.

8 Dataset and Performance Evaluation

To quantitatively demonstrate the effectiveness and efficiency of various image retrieval algorithms, it is indispensable to collect some benchmark datasets and define the evaluation metrics. In this section, we discuss the recent ground truth datasets and distractor datasets used in experimental study for image retrieval. Besides, we introduce the key evaluation indicators in CBIR, including accuracy, efficiency, and memory cost.

8.1 Recent Dataset for CBIR

Intuitively, the ground-truth dataset should be sufficient large so as to well demonstrate the scalability of image retrieval algorithms. However, considering the tedious labor in dataset collection, the existing ground-truth dataset are relatively small, but mixed with random million-scale distractor database for evaluation on scalability. The existing ground-truth datasets target on particular object/scene retrieval or partial-duplicate Web image retrieval. Generally, the ground-truth images contain a specific object or scene and may undergo various changes and be taken under different views or changes in illumination, scale, rotation, partial occlusion, compression rate, *etc*. Typical ground truth dataset for this task includes the UKBench dataset [10], the Oxford Building dataset [11], and the Holidays dataset [12], *etc*. MIR Flickr-1M and Flickr-1M are two different million-scale databases which are usually used as distractor to evaluate the scalability of image retrieval algorithms. For convenience of comparison and reference, we list the general information of those recent datasets popularly used in CBIR in Table 1. Some sample images from those datasets are shown in Fig. 3.

UKBench dataset It contains 10,200 images from 2,550 categories[9]. In each category, there are four images taken on the same scene or object from different views or illumination conditions. All the 10,200 images are taken as query and their retrieval performances are averaged.

Holidays dataset There are 1,491 images from 500 groups in the Holidays dataset[10]. Images in each group are taken on a scene or an object with various viewpoints. The first image in each group is selected as query for evaluation.

Oxford Building dataset (Oxford-5K) The Oxford Buildings Dataset[11] consists of 5062 images collected from Flickr[12] by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. Some junk images are mixed in it as distractor.

9. http://www.vis.uky.edu/∼stewe/ukbench/
10. http://lear.inrialpes.fr/people/jegou/data.php
11. http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/
12. http://www.flickr.com/

TABLE 1
General information of the popular retrieval datasets in CBIR. The "mixed" database type denotes that the corresponding dataset is a ground truth dataset mixed with distractor images.

Database Name	Database Type	Database Size	Query Number	Category Number	Resolution
UKBench	Ground Truth	10,200	10,200	2,550	640×480
Holidays	Ground Truth	1,491	500	500	1024×768
Oxford-5K	Mixed	6,053	55	11	1024×768
Paris	Mixed	6,412	500	12	1024×768
DupImage	Ground Truth	1,104	108	33	460×350 (average)
FlickrLogos-32	Mixed	8,240	500	32	1024×768
INSTRE	Ground Truth	28,543	N/A	200	1000×720 (average)
ZuBuD	Ground Truth	1,005	115	200	320×240
SMVS	Ground Truth	1,200	3,300	1,200	640×480
MIR Flickr-1M	Distractor	1,000,000	N/A	N/A	500×500
Flickr1M	Distractor	1,000,000	N/A	N/A	N/A

Paris dataset In the Paris dataset[13], there are 6,412 images, which are collected from Flickr by searching for 12 text queries of particular Paris landmarks. For this dataset, 500 query images are used for evaluation.

DupImage dataset This dataset contains 1,104 images from 33 groups[14]. Each group corresponds to a logo, a painting, or an artwork, such as KFC, American Gothic Painting, Mona Lisa, *etc*. 108 representative query images are selected from those groups for evaluation.

FlickrLogos-32 dataset This dataset[15] contains logo images of 32 different brands which are downloaded from Flickr. All logo images in this dataset have an approximately planar structure. The dataset is partitioned into three subsets for evaluation, *i.e.*, training set, validation set, and query set [220]. Of those 8,240 images in the dataset, 6,000 images contain no logos and works as distractors.

INSTRE As an instance-level benchmark dataset, the INSTRE dataset [16] contains two subsets, *i.e.*, INSTRE-S and INSTRE-M [221]. In the former subset, there are 23,070 images, each with a single label of 200 classes. The latter subset contains 5,473 images and each image contains two instances from 100 object categories.

ZuBuD dataset The basic dataset contains 1,005 images of 201 buildings in Zurich, with 5 views for each building[17]. Besides, there are additional 115 query images which are not included in the basic dataset. The resolution of those images are uniformly 320×240.

Stanford Mobile Visual Search (SMVS) Dataset This dataset[18] is targeted for mobile visual search and contains images taken by camera phone on products, CDs, books, outdoor landmarks, business cards, text documents, museum paintings and video clips. It is characterized by rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, and realistic ground-truth reference data [222]. In the dataset, there are 1,200 distinct categories. For each category, one reference image with resolution quality is collected for evaluation. There are 3,300 query images in total which are collected from heterogeneous low and high-end camera phones.

13. http://www.robots.ox.ac.uk/∼vgg/data/parisbuildings/
14. http://pan.baidu.com/s/1jGETFUm
15. http://www.multimedia-computing.de/flickrlogos/
16. http://vipl.ict.ac.cn/isia/instre/
17. http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html
18. http://purl.stanford.edu/rb470rw0983

MIR Flickr-1M This is a distractor dataset[19], with one million images randomly downloaded from Flickr and resized to be no larger than 500 by 500.

Flickr1M is another distractor database containing SIFT features[20] of one million images arbitrarily retrieved from Flickr. The original images in this database are not available.

8.2 Performance Evaluation for CBIR

In the design of a multimedia content-based retrieval system, there are three key indicators which should be carefully considered: accuracy, efficiency, and memory cost. Usually, a retrieval method contributes to improving at least one of those indicators with little sacrifice in the other indicators.

Accuracy To measure the retrieval quality quantitatively, the database images are categorized into difference relevance levels and the accuracy score is summarized based on the rank order of the database images. For different relevance levels, there are different accuracy metrics. Where there are only two relevance level, *i.e.*, relevant and irrelevant, average precision (AP) is widely used to evaluate the retrieval quality of a single query's retrieval results. AP takes consideration of both precision and recall. Precision denotes the fraction of retrieved (top k) images that are relevant while recall means fraction of relevant image that are retrieved (in the top k returned results). Generally, for a retrieval system, precision decreases as either the number of images retrieved increases or recall grows. AP averages the precision values from the rank positions where a relevant image was retrieved, as defined in Eq. 7. To summarize the retrieval quality over multiple query images, the mean average precision (mAP) is usually adopted, which average the average precision over all queries.

$$AP = \frac{\sum_{k=1}^{n} P(k) \cdot rel(k)}{R} \tag{7}$$

where R denotes the number of relevant results for the current query image, $P(k)$ denotes the precision of top k retrieval results, $rel(k)$ is a binary indicator function equalling 1 when the k-th retrieved result is relevant to the current query image and 0 otherwise, and n denotes the total number of retrieved results.

19. http://medialab.liacs.nl/mirflickr/mirflickr1m/
20. http://bigimbaz.inrialpes.fr/herve/siftgeo1M/

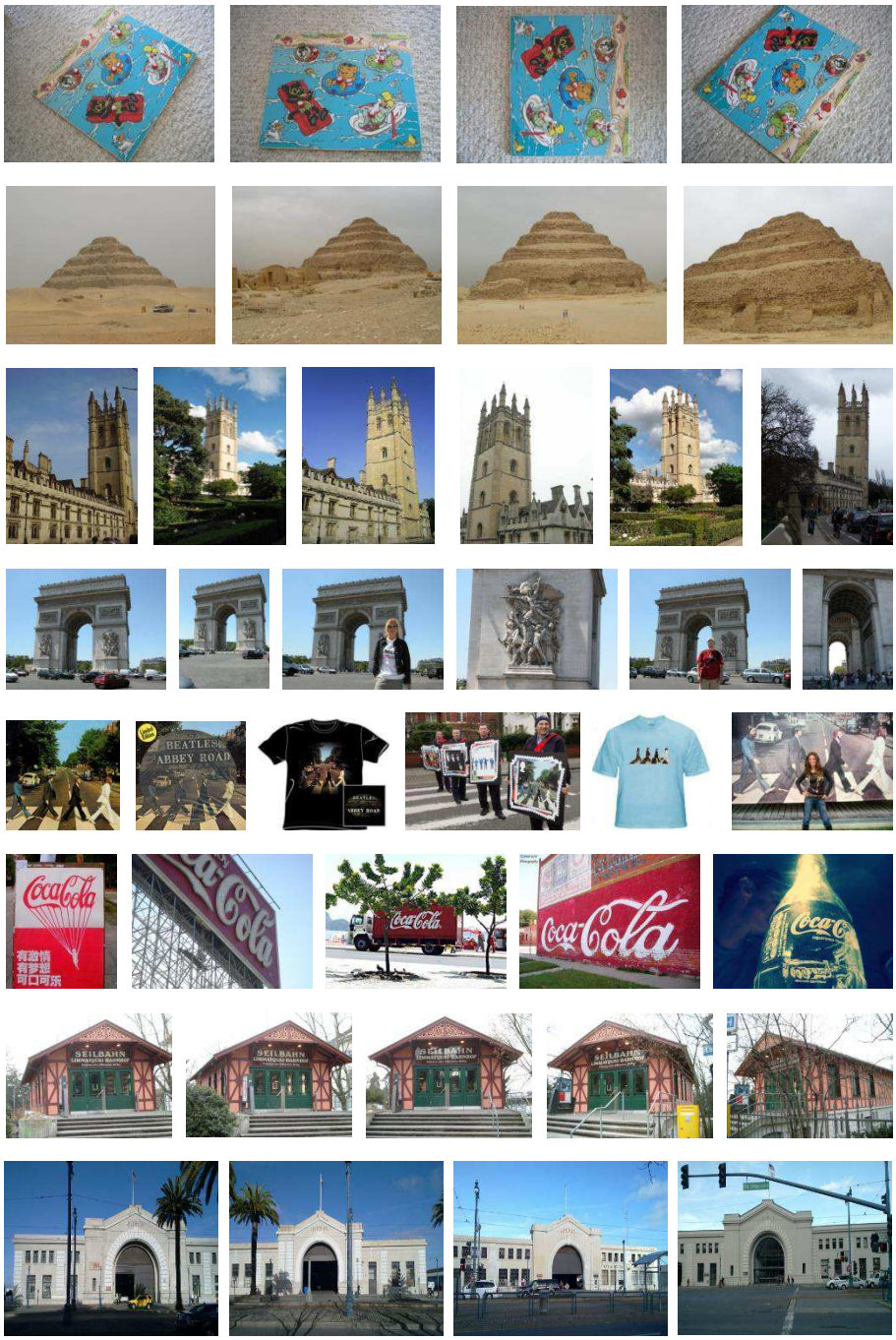

Fig. 3. Samples images of the existing datasets. First row: UKBench dataset; second row: Holidays dataset; third row: Oxford Building dataset; fourth row: DupImage dataset; fifth row: INSTRE dataset; sixth row: ZuBuD dataset; seventh row: SMVS dataset.

When there are multiple relevance levels, we can resort to normalized discounted cumulative gain (NDCG) metric defined in Eq. 8 to summarize the ranking results.

$$NDCG = \frac{1}{N}(r_1 + \sum_{k=2}^{n} \frac{f(r_k)}{log_2(k)}), \tag{8}$$

where n denotes the number of retrieved images, r_k denotes the relevance level, $f(\cdot)$ is function to tune the contribution of difference relevance levels, and N denotes the normalized term to ensure that the NDCG score for the ideal retrieved results is 100%. Popular definitions of $f(\cdot)$ include $f(x) = x$ and $f(x) = 2^x - 1$, with the latter to emphasize on retrieving highly relevant images.

Besides the above measures, some simple measures may be adopted for special dataset. In the public UKBench dataset, considering that there are four relevant images for all queries, the N-S score, *i.e.*, the average 4 times top-4 precision over the dataset, are used to measure the retrieval accuracy [10].

Computational Efficiency The efficiency of a retrieval system involves the time cost in visual vocabulary (codebook) construction, visual feature indexing, and image querying. The first two items are performed off-line, while

the last one is conducted on-line. Both the off-line and on-line processing is expected to be as fast as possible. Specially, the on-line querying is usually expected to be responded in real time.

Memory Cost In a multimedia content-based visual retrieval system, the memory cost usually refers to the memory usage in the on-line query stage. Generally, the memory is mainly spent on the quantizer and the index file of database, which need to be loaded into the main memory for on-line retrieval. Popular quantizer includes tree-based structure, such as hierarchical vocabulary tree, randomized forests, *etc*, which usually cost a few hundred mega-bytes memory for codebook containing million-scale visual words. In some binary code based quantization methods [36] [72], the quantizer is simple hash function with negligible memory overhead. For the index file, the memory cost is proportional to the indexed database size. When the database images are represented by local features and each local feature is indexed locally, the index file is proportional to the amount of indexed features and the memory cost per indexed feature.

9 FUTURE DIRECTIONS

Despite the extensive research efforts in the past decade, there is still sufficient space to further boost content based visual search. In the following, we will discuss several directions for future research, on which new advance shall be made in the next decade.

9.1 Ground-Truth Dataset Collection

In the multimedia and computer vision field, ground-truth datasets are motivated by the specific tasks. At the beginning of those dataset construction, they inspire researchers to update the performance records with their best efforts, leading to many classic ideas and algorithms to address the research problem. However, with the advance to address those datasets, the break-through of some algorithms may suffer from the over-fitting to the dataset. Meanwhile, with deeper understanding and clearer definition of the research problem, the limitation of existing datasets is revealed and new datasets are expected. For content-based image retrieval, we also expect better ground-truth dataset to be collected and released. Generally, the new ground-truth datasets shall be specific to eliminate the ambiguity of relevance of image content, such as logo datasets. Meanwhile, the scale of the dataset shall be sufficiently large so as to distinguish the problem of CBIR from image classification.

9.2 Intention Oriented Query Formation and Selection

Intention gap is the first and of the greatest challenge in content-based image retrieval. A simple query in the form of example, color map or sketch map is still insufficient in most time to reflect the user intention, consequently generating unsatisfactory retrieval results. Besides the traditional query formations, assistance from user to specify the concrete expectation will greatly alleviate the difficulty of the following image retrieval process. Considering that the end-users may be reluctant to involve much in the query formation, it is still possible to design convenient query formation interface to reduce the user involvement as much as possible. For instance, it is easy for a user to specify the region of interest in an example image for retrieval, or indicate the expected results are partial-duplicates or just similar in spatial color and texture. It is also possible to predict the potential intentions based on the initial query and make confirmation with end-user. In all, rather than passively induce the intension behind the query, it is beneficial to actively involve end-user in the retrieval process.

In image retrieval, the search performance is significantly impacted by the quality of the query. How to select a suitable query towards the optimal retrieval is a nontrivial issue. The query quality is related with many factors, including resolution, noise pollution, affine distortion, background clutter, *etc*. In the scenario of mobile search, the query can be selected by guiding the end user to retake better photos. In the server end, automatic retrieval quality assessment methods [223] [224] can be designed to select potential candidate from the initial retrieval results of high precision.

9.3 Deep Learning in CBIR

Despite the advance in content-based visual retrieval, there is still significant gap towards semantic-aware retrieval from visual content. This is essentially due to the fact that current image representation schemes are hand-crafted and insufficient to capture the semantics. Due to the tremendous diversity and quantity in multimedia visual data, most existing methods are un-supervised. To proceed towards semantic-aware retrieval, scalable scalable supervised or semi-supervised learning are promising to learn semantic-aware representation so as to boost the content-based retrieval quality. The success of deep learning in large-scale visual recognition [99] [96] [95] [225] has already demonstrated such potential.

To adapt those existing deep learning techniques to CBIR, there are several non-trivial issues that deserve research efforts. Firstly, the learned image representation with deep learning shall be flexible and robust to various common changes and transformations, such as rotation and scaling. Since the existing deep learning relies on the convolutional operation with anisotropic filters to convolve images, the resulted feature maps are sensitive to large translation, rotation, and scaling changes. It is still an open problem as whether that can solved by simply including more training samples with diverse transformations. Secondly, since computational efficiency and memory overhead are emphasized in particular in CBIR, it would be beneficial to consider those constraints in the structure design of deep learning networks. For instance, both compact binary semantic hashing codes [59] [65] and very sparse semantic vector representations are desired to represent images, since the latter are efficient in both distance computing and memory storing while the latter is well adapted to the inverted index structure.

9.4 Unsupervised Database Mining

In traditional content-based image retrieval algorithms and systems, the database images are processed independently without considering their potential relevance context information. This is primarily due to the fact that, there is

usually no label information for the database images and the potential category number is unlimited. Those constraints limit the application of sophisticated supervised learning algorithms in CBIR. However, as long as the database is large, it is likely that there exist some subsets of images and images in each sub-set are relevant to each other images. Therefore, it is feasible to explore the database images with some unsupervised techniques to uncover those sub-sets in the off-line processing stage. If we regard each database image as a node and the relevance level between images as edge to link images, the whole image database can be represented as a large graph. Then, the sub-sets mining problem can be formulated as a sub-graph discovery problem. On the other hand, in practice, new images may be incrementally included into the graph, which casts challenge to dynamically uncover those sub-graphs on the fly. The mining results in the off-line stage will be beneficial for the on-line query to yield better retrieval results.

9.5 Cross-modal Retrieval

In the above discussion of this survey, we focus on the visual content for image retrieval. However, besides the visual features, there are other very useful clues, such as the textual information around images in Web pages, the click log of users when using the search engines, the speech information in videos, *etc*. Those multi-modal clues are complementary to each to collaboratively identify the visual content of images and videos. Therefore, it would be beneficial to explore cross-modal retrieval and fuse those multi-modal features with different models. With multi-modal representation, there are still many open search topics in terms of collaborative quantization, indexing, search re-ranking, *etc*.

9.6 End-to-End Retrieval Framework

As discussed in the above sections, the retrieval framework is involved with multiple modules, including feature extraction, codebook learning, feature quantization, feature quantization, image indexing, *etc*. Those modules are individually designed and independently optimized for the retrieval task. On the other hand, if we investigate the structure of the convolutional neural network (CNN) in deep learning, we can find a very close analogy between the BoW model and the CNN model. The convolutional filters used in the CNN model works in a similar way as the codewords of the codebook in the BoW model. The convolution results between the image patch and the convolution filter are essentially the soft quantization results, with the max-pooling operation similar to the local aggregation in the BoW model. As long as the learned feature vector is sparse, we can also adopt the inverted index structure to efficiently index the image database. Different from the BoW model, the above modules in the CNN model are collaboratively optimized for the task of image classification. Based on the above analogy, similarly, we may also resort to an end-to-end paradigm to design a framework that takes images as input and outputs the index-oriented features directly, with the traditional key retrieval-related modules implicitly and collaboratively optimized.

9.7 Social Media Mining with CBIR

Different from the traditional unstructured Web media, the emerging social media in recent years have been characterized by community based personalized content creation, sharing, and interaction. There are many successful prominent platforms of social media, such as Facebook, Twitter, Wikipedia, LinkedIn, Pinterest, *etc*. The social media is enriched with tremendous information which dynamically reflects the social and cultural background and trend of the community. Besides, it also reveals the personal affection and behavior characteristics. As an important media of the user-created content, the visual data can be used as an entry point with the content-based image retrieval technique to uncover and understand the underlying community structure. It would be beneficial to understand the behavior of individual users and conduct recommendation of products and services to users. Moreover, it is feasible to analyze the sentiment of crowd for supervision and forewarning.

9.8 Open Grand Challenge

Due to the difference in deployment structure and availability of data, the research on content based image retrieval in the academia suffers a gap from the real application in industry. To bridge this gap, it is beneficial to initiate some open grand challenges from the industry and involve the researchers in the academia to investigate the key difficulties in real scenarios. In the past five years, there are some limited open grand challenge, such as the Microsoft Image Grand Challenge on Image Retrieval[21] and Alibaba Large-Scale Image Search Challenge[22]. In the future, we would expect many more such grand challenges. The open grand challenge will only only advance the research progress in the academia, but also benefit the industry with more and better practical and feasible solutions to the real-world challenges.

10 CONCLUSIONS

In this paper, we have investigated the advance on content-based image retrieval in recent years. We focus on the five key modules of the general framework, *i.e.*, query formation, image representation, image indexing, retrieval scoring, and search re-ranking. For each component, we have discussed the key problems and categorized a variety of representative strategies and methods. Further, we have summarized eight potential directions that may boost the advance of content based image retrieval in the near future.

REFERENCES

[1] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, 1998.

[2] A. Alzubi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 20–54, 2015.

21. http://acmmm13.org/submissions/call-for-multimedia-grand-challenge-solutions/msr-bing-grand-challenge-on-image-retrieval-scientific-track

22. http://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100069.5678.1.SmufkG&raceId=231510&_lang=en_US

[3] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 14, 2016.

[4] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3864–3872.

[5] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[7] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[8] D. G. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 1470–1477.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2161–2168.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[12] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision*, 2008, pp. 304–317.

[13] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Large scale image search with geometric coding," in *ACM International Conference on Multimedia*, 2011, pp. 1349–1352.

[14] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[16] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *British Machine Vision Conference*, vol. 3, 2008, p. 4.

[17] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 25–32.

[18] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *ACM International Conference on Multimedia*, 2010, pp. 511–520.

[19] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 889–896.

[20] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 809–816.

[21] X. Zhang, L. Zhang, and H.-Y. Shum, "Qsrank: Query-sensitive hash code ranking for efficient ϵ-neighbor search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2058–2065.

[22] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3005–3012.

[23] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2911–2918.

[24] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *European Conference on Computer Vision (ECCV)*, 2012.

[25] Q. Tian, S. Zhang, W. Zhou, R. Ji, B. Ni, and N. Sebe, "Building descriptive and discriminative visual codebook for large-scale image applications," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 441–477, 2011.

[26] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Large scale partial-duplicate image retrieval with bi-space quantization and geometric consistency," in *IEEE International Conference Acoustics Speech and Signal Processing*, 2010, pp. 2394–2397.

[27] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *ACM International Conference on Multimedia*, 2009, pp. 75–84.

[28] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, "Building contextual visual vocabulary for large-scale image applications," in *ACM International Conference on Multimedia*, 2010, pp. 501–510.

[29] W. Zhou, Q. Tian, Y. Lu, L. Yang, and H. Li, "Latent visual context learning for web image applications," *Pattern Recognition*, vol. 44, no. 10, pp. 2263–2273, 2011.

[30] G. Tolias, Y. Avrithis, and H. Jgou, "To aggregate or not to aggregate: selective match kernels for image search," in *International Conference on Computer Vision (ICCV)*, 2013.

[31] L. Zhang and Y. Rui, "Image searchłfrom thousands to billions in 20 years," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1s, p. 36, 2013.

[32] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang, "Intentsearch: Capturing user intention for one-click internet image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 7, pp. 1342–1353, 2012.

[33] B. Moghaddam, Q. Tian, N. Lesh, C. Shen, and T. S. Huang, "Visualization and user-modeling for browsing personal photo libraries," *International Journal of Computer Vision (IJCV)*, vol. 56, no. 1-2, pp. 109–130, 2004.

[34] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[35] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[36] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *ACM International Conference on Multimedia*, 2012, pp. 169–178.

[37] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Mindfinder: interactive sketch-based image search on millions of images," in *ACM International Conference on Multimedia (MM)*, 2010, pp. 1605–1608.

[38] C. Xiao, C. Wang, L. Zhang, and L. Zhang, "Sketch-based image retrieval via shape words," in *ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2015, pp. 571–574.

[39] P. Sousa and M. J. Fonseca, "Sketch-based retrieval of drawings using spatial proximity," *Journal of Visual Languages & Computing*, vol. 21, no. 2, pp. 69–80, 2010.

[40] M. J. Fonseca, A. Ferreira, and J. A. Jorge, "Sketch-based retrieval of complex drawings using hierarchical topology and geometry," *Computer-Aided Design*, vol. 41, no. 12, pp. 1067–1081, 2009.

[41] S. Liang and Z. Sun, "Sketch retrieval and relevance feedback with biased svm classification," *Pattern Recognition Letters*, vol. 29, no. 12, pp. 1733–1741, 2008.

[42] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 761–768.

[43] J. Wang and X.-S. Hua, "Interactive image search by color map," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 1, p. 12, 2011.

[44] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Image search by concept map," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2010, pp. 275–282.

[45] ——, "Interactive image search by 2d semantic map," in *International Conference on World Wide Web (WWW)*. ACM, 2010, pp. 1321–1324.

[46] T. Lan, W. Yang, Y. Wang, and G. Mori, "Image retrieval with structured object queries using latent ranking svm," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 129–142.

[47] G. Kim, S. Moon, and L. Sigal, "Ranking and retrieval of image sequences from multiple paragraph queries," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1993–2001.

[48] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *ACM International Conference on Multimedia*. ACM, 2011, pp. 1437–1440.

[49] J. Xie, Y. Fang, F. Zhu, and E. Wong, "Deepshape: Deep learned shape descriptor for 3d shape matching and retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1275–1283.

[50] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1875–1883.

[51] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. Jan Latecki, "Gift: A real-time and scalable 3d shape search engine," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5023–5032.

[52] M. Park, J. S. Jin, and L. S. Wilson, "Fast content-based image retrieval using quasi-gabor filter and reduction of image feature dimension," in *IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE, 2002, pp. 178–182.

[53] X.-Y. Wang, B.-B. Zhang, and H.-Y. Yang, "Content-based image retrieval by integrating color and texture features," *Multimedia Tools and Applications (MTA)*, vol. 68, no. 3, pp. 545–569, 2014.

[54] B. Wang, Z. Li, M. Li, and W.-Y. Ma, "Large-scale duplicate detection for web image search," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2006, pp. 353–356.

[55] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, no. 2, pp. 300–312, 2007.

[56] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *International Conference on Computer Vision*, 2009, pp. 2130–2137.

[57] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1753–1760.

[58] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[59] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[60] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[61] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[62] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[63] H. Xie, K. Gao, Y. Zhang, S. Tang, J. Li, and Y. Liu, "Efficient feature detection and effective post-verification for large scale near-duplicate image search," *IEEE Transactions on Multimedia (TMM)*, vol. 13, no. 6, pp. 1319–1332, 2011.

[64] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2010.

[65] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *ESANN*. Citeseer, 2011.

[66] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "A multi-sample, multi-tree approach to bag-of-words image representation for image retrieval," in *IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 1992–1999.

[67] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417.

[68] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[69] W. Zhou, H. Li, R. Hong, Y. Lu, and Q. Tian, "BSIFT: towards data-independent codebook for large scale image search," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 3, pp. 967–979, 2015.

[70] Z. Liu, H. Li, L. Zhang, W. Zhou, and Q. Tian, "Cross-indexing of binary SIFT codes for large-scale image search," *IEEE Transactions on Image Processing (TIP)*, 2014.

[71] G. Yu and J.-M. Morel, "Asift: an algorithm for fully affine invariant comparison," *Image Processing On Line*, vol. 2011, 2011.

[72] W. Dong, Z. Wang, M. Charikar, and K. Li, "High-confidence near-duplicate image detection," in *ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2012, p. 1.

[73] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: binary robust independent elementary features," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 778–792.

[74] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *International Conference on Computer Vision*, 2011, pp. 2564–2571.

[75] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 510–517.

[76] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: binary robust invariant scalable keypoints," in *International Conference on Computer Vision*, 2011, pp. 2548–2555.

[77] S. Zhang, Q. Tian, Q. Huang, W. Gao, and Y. Rui, "USB: Ultra-short binary descriptor for fast visual matching and retrieval," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 8, pp. 3671–3683, 2014.

[78] S. Madeo and M. Bober, "Fast, compact and discriminative: Evaluation of binary descriptors for mobile applications," *IEEE Transactions on Multimedia*, 2016.

[79] S. Zhang, Q. Tian, K. Lu, Q. Huang, and W. Gao, "Edge-SIFT: Discriminative binary descriptor for scalable partial-duplicate mobile search," *IEEE Transactions on Image Processing*, 2013.

[80] K. E. Van De Sande, T. Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1582–1596, 2010.

[81] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 745–752.

[82] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *ACM International Conference on Multimedia (MM)*. ACM, 2014, pp. 1025–1028.

[83] R. Tao, A. W. Smeulders, and S.-F. Chang, "Attributes and categories for generic instance search from one example," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 177–186.

[84] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1778–1785.

[85] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3306–3313.

[86] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 776–789.

[87] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 785–792.

[88] J. Cai, Z.-J. Zha, M. Wang, S. Zhang, and Q. Tian, "An attribute-assisted reranking model for web image search," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 1, pp. 261–272, 2015.

[89] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware co-indexing for image retrieval," in *IEEE International Conference on Computer Vis*, 2013.

[90] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," in *British Machine Vision Conference (BMVC)*, 2014.

[91] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[92] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[93] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *ACM International Conference on Image and Video Retrieval*, 2007, pp. 17–24.

[94] R. Lienhart and M. Slaney, "pLSA on large scale image databases," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–1217.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[96] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[97] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[98] E. Hörster and R. Lienhart, "Deep networks for image retrieval on large-scale databases," in *ACM International Conference on Multimedia*. ACM, 2008, pp. 643–646.

[99] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[100] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[101] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *ACM International Conference on Multimedia (MM)*. ACM, 2014, pp. 157–166.

[102] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Visual instance retrieval with deep convolutional networks," *arXiv preprint arXiv:1412.6574*, 2014.

[103] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2015.

[104] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2015.

[105] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision (IJCV)*, vol. 104, no. 2, pp. 154–171, 2013.

[106] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2189–2202, 2012.

[107] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[108] S. Sun, W. Zhou, Q. Tian, and H. Li, "Scalable object retrieval with compact image representation from generic object regions," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 2, p. 29, 2015.

[109] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *International Conference on Learning and Representation (ICLR)*, 2016.

[110] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *European Conference on Computer Vision (ECCV)*, 2016.

[111] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[112] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 584–599.

[113] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid, "Local convolutional features with unsupervised training for image retrieval," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 91–99.

[114] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2156–2162.

[115] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," *arXiv preprint arXiv:1504.03410*, 2015.

[116] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[117] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3384–3391.

[118] F. Li, W. Tong, R. Jin, A. K. Jain, and J.-E. Lee, "An efficient key point quantization algorithm for large scale image retrieval," in *ACM workshop on Large-scale Multimedia Retrieval and Mining*. ACM, 2009, pp. 89–96.

[119] L. Chu, S. Wang, Y. Zhang, S. Jiang, and Q. Huang, "Graph-density-based visual word vocabulary for image retrieval," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[120] W. Dong, Z. Wang, M. Charikar, and K. Li, "Efficiently matching sets of features with random histograms," in *ACM International Conference on Multimedia (MM)*. ACM, 2008, pp. 179–188.

[121] W. Zhou, M. Yang, H. Li, X. Wang, Y. Lin, and Q. Tian, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 601–611, 2014.

[122] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 9, pp. 2664–2677, 2011.

[123] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *International Conference on Computer Vision*, 2011, pp. 209–216.

[124] Z. Liu, H. Li, W. Zhou, and Q. Tian, "Embedding spatial context information into inverted file for large-scale image retrieval," in *ACM International Conference on Multimedia*, 2012, pp. 199–208.

[125] Z. Liu, H. Li, W. Zhou, R. Zhao, and Q. Tian, "Contextual hashing for large-scale image search," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 4, pp. 1606–1614, 2014.

[126] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 17–24.

[127] D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 4, pp. 415–423, 1998.

[128] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[129] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-bag-of-features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3352–3359.

[130] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multireference reranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1991–2001, 2011.

[131] J. L. Bentley, "K-d trees for semidynamic point sets," in *Annual Symp. Computational Geometry*, 1990, pp. 187–197.

[132] C. Silpa-Anan and R. Hartley, "Localization using an image map," in *Australian Conference on Robotics and Automation*, 2004.

[133] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, 2014.

[134] W. Zhou, M. Yang, X. Wang, H. Li, Y. Lin, and Q. Tian, "Scalable feature matching by dual cascaded scalar quantization for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 1, pp. 159–171, 2016.

[135] M. Jain, H. Jégou, and P. Gros, "Asymmetric hamming embedding: taking the best of our bits for large scale image search," in *ACM International Conference on Multimedia*, 2011, pp. 1441–1444.

[136] W. Zhou, H. Li, Y. Lu, M. Wang, and Q. Tian, "Visual word expansion and BSIFT verification for large-scale image search," *Multimedia Systems*, vol. 21, no. 3, pp. 245–254, 2013.

[137] Y. Xia, K. He, F. Wen, and J. Sun, "Joint inverted indexing," in *International Conference on Computer Vision*, 2013.

[138] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening," in *European Conference on Computer Vision*, 2012, pp. 774–787.

[139] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[140] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Annual ACM Symposium Theory of Computing*. ACM, 1998, pp. 604–613.

[141] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *IEEE Symposium Foundations of Computer Science*, 2006, pp. 459–468.

[142] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe lsh: efficient indexing for high-dimensional similarity search," in *International Conference Very Large Data Bases*, 2007, pp. 950–961.

[143] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3424–3431.

[144] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 817–824.

[145] D. Aiger, E. Kokiopoulou, and E. Rivlin, "Random grids: Fast approximate nearest neighbors and range searching for image search," in *International Conference on Computer Vision*, 2013.

[146] M. Iwamura, T. Sato, and K. Kise, "What is the most efficient way to select nearest neighbor candidates for fast approximate nearest neighbor search?" in *International Conference on Computer Vision*, 2013.

[147] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for large scale indexing," in *ACM International Conference on Multimedia (MM)*. ACM, 2012, pp. 179–188.

[148] M. Wang, W. Zhou, Q. Tian, Z. Zha, and H. Li, "Linear distance preserving pseudo-supervised and unsupervised hashing," in *ACM International Conference on Multimedia (MM)*. ACM, 2016, pp. 1257–1266.

[149] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization for approximate nearest neighbor search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[150] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[151] R. Arandjelovic and A. Zisserman, "All about vlad," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 1578–1585.

[152] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "A comprehensive study over vlad and product quantizationin for large-scale image retrieval," *IEEE Transactions on Multimedia (TMM)*, 2014.

[153] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 3310–3317.

[154] Z. Gao, J. Xue, W. Zhou, S. Pang, and Q. Tian, "Fast democratic aggregation and query fusion for image search," in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2015.

[155] T. Ge, Q. Ke, and J. Sun, "Sparse-coded features for image retrieval." British Machine Vision Conference (BMVC), 2013.

[156] Z. Liu, H. Li, W. Zhou, T. Rui, and Q. Tian, "Uniforming residual vector distribution for distinctive image representation," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2015.

[157] Z. Liu, H. Li, W. Zhou, and Q. Tian, "Uniting keypoints: Local visual information fusion for large scale image search," *IEEE Transactions on Multimedia (TMM)*, 2015.

[158] T. Jaakkola and D. Haussler, "Exploring generative model in discriminative classifiers," in *Advances in Neural Information Processing Systems (NIPS)*, 1998.

[159] T. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in Neural Information Processing Systems (NIPS)*, pp. 487–493, 1999.

[160] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 105, no. 3, pp. 222–245, 2013.

[161] L.-Y. Duan, F. Gao, J. Chen, J. Lin, and T. Huang, "Compact descriptors for mobile visual search and mpeg cdvs standardization," in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013, pp. 885–888.

[162] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 392–407.

[163] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1269–1277.

[164] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York., 1999, vol. 463.

[165] J. Cai, Q. Liu, F. Chen, D. Joshi, and Q. Tian, "Scalable image search with multiple index tables," in *International Conference on Multimedia Retrieval (ICMR)*. ACM, 2014, p. 407.

[166] L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large-scale image retrieval," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 8, pp. 3368–3380, 2014.

[167] X. Zhang, Z. Li, L. Zhang, W.-Y. Ma, and H.-Y. Shum, "Efficient indexing for large scale visual search," in *IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 1103–1110.

[168] C. Silpa-Anan and R. Hartley, "Optimised kd-trees for fast image descriptor matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[169] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Fast image retrieval: Query pruning and early termination," *IEEE Transactions on Multimedia (TMM)*, vol. 17, no. 5, pp. 648–659, 2015.

[170] R. Ji, L.-Y. Duan, J. Chen, L. Xie, H. Yao, and W. Gao, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Transactions on Multimedia (TMM)*, vol. 15, no. 1, pp. 153–166, 2013.

[171] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2957–2964.

[172] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Transactions on Image Processing (TPI)*, vol. 24, no. 9, pp. 2827–2840, 2015.

[173] L. Wu, K. Zhao, H. Lu, Z. Wei, and B. Lu, "Distance preserving marginal hashing for image retrieval," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.

[174] K. Jiang, Q. Que, and B. Kulis, "Revisiting kernelized locality-sensitive hashing for improved large-scale image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4933–4941.

[175] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2064–2072.

[176] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Annual Symposium on Computational Geometry*. ACM, 2004, pp. 253–262.

[177] Y. Avrithis, G. Tolias, and Y. Kalantidis, "Feature map hashing: sub-linear indexing of appearance and global geometry," in *ACM International Conference on Multimedia (MM)*. ACM, 2010, pp. 231–240.

[178] G. Tolias, Y. Kalantidis, Y. Avrithis, and S. Kollias, "Towards large-scale geometry indexing by feature selection," *Computer Vision and Image Understanding*, vol. 120, pp. 31–45, 2014.

[179] H. Jégou, M. Douze, and C. Schmid, "Packing bag-of-features," in *International Conference on Computer Vision*, 2009, pp. 2357–2364.

[180] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2007, pp. 549–556.

[181] Z. Lin and J. Brandt, "A local bag-of-features model for large-scale object retrieval," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 294–308.

[182] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek, "Accurate image search using the contextual dissimilarity measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 2–11, 2010.

[183] D. Qin, C. Wengert, and L. Van Gool, "Query adaptive similarity for large scale object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 1610–1617.

[184] M. Donoser and H. Bischof, "Diffusion processes for retrieval revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1320–1327.

[185] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Lp-norm IDF for large scale image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[186] L. Zheng, S. Wang, and Q. Tian, "Lp-norm IDF for scalable image retrieval," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3604–3617, 2014.

[187] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3013–3020.

[188] H. Xie, K. Gao, Y. Zhang, J. Li, and Y. Liu, "Pairwise weak geometric consistency for large scale image search," in *ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2011, p. 42.

[189] S. M. Katz, "Distribution of content words and phrases in text and language modelling," *Natural Language Engineering*, vol. 2, no. 01, pp. 15–59, 1996.

[190] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1169–1176.

[191] M. Shi, Y. Avrithis, and H. Jégou, "Early burst detection for memory-efficient image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[192] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1056–1069, 2016.

[193] F. Yang, B. Matei, and L. S. Davis, "Re-ranking by multi-feature fusion with diffusion for image retrieval," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2015, pp. 572–579.

[194] X. Li, M. Larson, and A. Hanjalic, "Pairwise geometric matching for large-scale object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5153–5161.

[195] Y.-H. Kuo, K.-T. Chen, C.-H. Chiang, and W. H. Hsu, "Query expansion for hash-based image object retrieval," in *ACM International Conference on Multimedia*, 2009, pp. 65–74.

[196] O. Chum and J. Matas, "Matching with prosac-progressive sample consensus," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 220–226.

[197] G. Tolias and Y. Avrithis, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[198] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[199] Y. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *International Journal of Computer Vision*, vol. 107, no. 1, pp. 1–19, 2014.

[200] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2. IEEE, 2005, pp. 1458–1465.

[201] W. Zhou, H. Li, Y. Lu, and Q. Tian, "SIFT match verification by geometric coding for large-scale partial-duplicate web image search," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1, p. 4, 2013.

[202] L. Chu, S. Jiang, S. Wang, Y. Zhang, and Q. Huang, "Robust spatial consistency graph model for partial duplicate image retrieval," *IEEE Transactions on Multimedia (TMM)*, vol. 15, no. 8, pp. 1982–1996, 2013.

[203] L. Xie, Q. Tian, W. Zhou, and B. Zhang, "Fast and accurate near-duplicate image search with affinity propagation on the imageweb," *Computer Vision and Image Understanding*, vol. 124, pp. 31–41, 2014.

[204] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[205] L. Xie, Q. Tian, W. Zhou, and B. Zhang, "Heterogeneous graph propagation for large-scale web image search," *IEEE Transactions on Image Processing (TIP)*, 2015.

[206] H. Xie, Y. Zhang, J. Tan, L. Guo, and J. Li, "Contextual query expansion for image retrieval," *IEEE Transactions on Multimedia (TMM)*, vol. 16, no. 4, pp. 1104–1114, 2014.

[207] D. Tao and X. Tang, "Random sampling based svm for relevance feedback image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[208] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28, no. 7, pp. 1088–1099, 2006.

[209] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–7.

[210] M. Arevalillo-Herráez and F. J. Ferri, "An improved distance-based relevance feedback strategy for image retrieval," *Image and Vision Computing (IVC)*, vol. 31, no. 10, pp. 704–713, 2013.

[211] E. Rabinovich, O. Rom, and O. Kurland, "Utilizing relevance feedback in fusion-based retrieval," in *International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*. ACM, 2014, pp. 313–322.

[212] X.-Y. Wang, Y.-W. Li, H.-Y. Yang, and J.-W. Chen, "An image retrieval scheme with relevance feedback using feature reconstruction and svm reclassification," *Neurocomputing*, vol. 127, pp. 214–230, 2014.

[213] K. Tieu and P. Viola, "Boosting image retrieval," *International Journal of Computer Vision (IJCV)*, vol. 56, no. 1-2, pp. 17–36, 2004.

[214] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[215] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia systems*, vol. 8, no. 6, pp. 536–544, 2003.

[216] P. B. Patil and M. B. Kokare, "Relevance feedback in content based image retrieval: A review." *Journal of Applied Computer Science & Mathematics*, no. 10, 2011.

[217] R. Fagin, R. Kumar, and D. Sivakumar, "Efficient similarity search and classification via rank aggregation," in *ACM SIGMOD International Conference on Management of Data*. ACM, 2003, pp. 301–312.

[218] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[219] G. Ye, D. Liu, I.-H. Jhuo, S.-F. Chang *et al.*, "Robust late fusion with rank minimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3021–3028.

[220] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol, "Scalable logo recognition in real-world images," in *ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2011, p. 25.

[221] S. Wang and S. Jiang, "Instre: a new benchmark for instance-level object retrieval and recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 3, p. 37, 2015.

[222] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach *et al.*, "The stanford mobile visual search data set," in *ACM conference on Multimedia Systems*. ACM, 2011, pp. 117–122.

[223] X. Tian, Y. Lu, L. Yang, and Q. Tian, "Learning to judge image search results," in *ACM International Conference on Multimedia (MM)*. ACM, 2011, pp. 363–372.

[224] X. Tian, Q. Jia, and T. Mei, "Query difficulty estimation for image search with query reconstruction error," *IEEE Transactions on Multimedia (TMM)*, vol. 17, no. 1, pp. 79–91, 2015.

[225] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 346–361.