Image Search—From Thousands to Billions in 20 Years

LEI ZHANG and YONG RUI, Microsoft Research Asia

This article presents a comprehensive review and analysis on image search in the past 20 years, emphasizing the challenges and opportunities brought by the astonishing increase of dataset scales from thousands to billions in the same time period, which was witnessed first-hand by the authors as active participants in this research area. Starting with a retrospective review of three stages of image search in the history, the article highlights major breakthroughs around the year 2000 in image search features, indexing methods, and commercial systems, which marked the transition from stage two to stage three. Subsequent sections describe the image search research from four important aspects: system framework, feature extraction and image representation, indexing, and big data's potential. Based on the review, the concluding section discusses open research challenges and suggests future research directions in effective visual representation, image knowledge base construction, implicit user feedback and crowdsourcing, mobile image search, and creative multimedia interfaces.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms: Algorithms, Documentation, Performance

Additional Key Words and Phrases: Review, image retrieval, Web image search, content-based, visual representation, image feature, global feature, local feature, indexing, big data, image knowledge base

ACM Reference Format:

Zhang, L. and Rui, Y. 2013. Image search—from thousands to billions in 20 years. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1s, Article 36 (October 2013), 20 pages.
DOI: http://dx.doi.org/10.1145/2490823

1. INTRODUCTION

With the emergence of Web browsers 20 years ago, digital content has been generated and published in an explosive way. Digital contents are inherent multimedia: texts, images, graphics, video and audio clips, etc. How to effectively and efficiently retrieve relevant digital multimedia content is therefore of great importance. Multimedia search is the research area that pushes the boundary of the best theory, framework, and tools for searching multimedia content. In this article, we focus on image search, one of the most important areas in multimedia search.

We focus on images because vision is the most important channel for human beings to communicate with the environment and learn new knowledge. According to the American Optometric Association [AOA 2006], approximately 80% of the learning a child does occurs through his or her eyes. The eyes must see clearly and with accurate focus control. The brain must interpret the visual image from its

Authors' addresses: L. Zhang (corresponding author) and Y. Rui, Microsoft Research Asia, No. 5 Danling Street, Beijing 100080, P.R. China; email: {leizhang, yongrui}@microsoft.com.

DOI: http://dx.doi.org/10.1145/2490823

background, make assumptions as to its figure, and integrate the information gathered from peripheral vision and from other senses.

Human vision is far more superior to machine vision, but computers are much faster than human brains in computation. This raises an interesting research problem, that is, whether computers can mimic human vision functions and perform time-consuming tasks faster than human beings. Among such tasks, we particularly expect computers can help us search and organize a large collection of images. This function is increasingly desirable due to the explosive growth of both personal photos and web images. As a result, there has been a strong need demanding efficient image search and management.

According to the datasets and technologies used in image search, we can divide the timeline of the research in image search into three stages: *the text-based stage* (1970–1990), *the content-based stage* (1990–2000) and *the web-based stage* (2000–present). The key focus of this article is to review the research of image search in the past 20 years. But for the sake of a complete history, we will also briefly touch the first stage in 1970–1990.

1.1 The Text-Based Stage (1970–1990)

Early work of image retrieval is mainly text-based, where images were first annotated with text and then searched using a text-based approach from traditional database management systems (DBMS). However, since automatically generating descriptive texts for a wide spectrum of images is not feasible, most text-based image retrieval systems require manual annotation of images, which is a tedious and expensive task for large-scale image databases, making the traditional text-based approaches nonscalable.

In 1979, a conference on database techniques for pictorial applications was held in Florence [Blaser 1979] to connect the researchers working separately on databases and pictorial applications. Since then, researchers have been paying attention to the problem of pictorial data management due to its large application potential. Because DBMS in 1970s were designed primarily for efficient storage, retrieval, and manipulation of alphanumeric data, major focuses were to introduce simple image processing techniques for the efficient retrieval and manipulation of pictorial information from pictorial databases [Chang and Fu 1980a; 1980b; Chang and Kunil 1981; Chang et al. 1988]. Because only simple pictorial datasets were handled, the applications mainly reside in vertical domains, such as computer-aided design (CAD), geographic information system (GIS), and remote sensing.

1.2 The Content-Based Stage (1990–2000)

In the early 1990s, the dramatically increased number of images produced by various digital devices and commercial applications posed a great demand on retrieving and managing images based on their visual content. The early work based on textual annotation and simple image processing techniques cannot fulfil such a complex need especially for large-scale natural image collections. In 1992, the National Science Foundation organized a workshop on visual information management systems [Jain 1993] to identify new directions in image database management systems. Since then, many researchers from the communities of computer vision, database management, human-computer interface, and information retrieval have been attracted to this field. As a result, many techniques of visual information extraction, browsing, user query and interaction, and indexing have been developed, and a large number of academic and commercial image retrieval systems have been built.

The key hindrance for the research in 1990s (and nowadays) is the well-known semantic gap between low-level image features and high-level semantic concepts. That is, users seek semantic similarity, but the database can only provide similarity by data processing. Smeulders et al. [2000] pointed out that content-based image retrieval does not rely on describing the content of the image in its entirety, and it

may be sufficient that a retrieval system presents similar images, similar in some user-defined sense. Indeed, a few systems were successfully developed and demonstrated good performance for some real applications. However, in many cases, the retrieval accuracy is still far from satisfactory.

To bridge the semantic gap, researchers resorted to users' feedbacks in an iterative way. This direction is called *relevance feedback* [Rui et al. 1998]. The major limitation of relevance feedback is that the user interface is not natural and users can be reluctant to provide explicit feedbacks. A comprehensive survey can be found in Zhou and Huang [2003].

1.3 The Web-Based Stage (2000–Present)

Since 2000, the advances in the Internet and digital imaging devices have significantly increased the number of images on the Web. In 2001, Google launched its image search engine, offering access to 250 million images. Although the index is purely based on the keywords extracted from the surrounding texts of Web images, image search results are often surprisingly good. This is because Web images usually have rich metadata, such as filename, URL, and surrounding text, which can be used as semantic-level descriptions for indexing and searching.

Motivated by the success of commercial Web image search engines, researchers started to look into large-scale Web images by studying a series of research problems, for example, (1) how to automatically generate textual keywords to annotate images? (2) How to build efficient index to scale up an image retrieval system? (3) How to implicitly collect users' feedbacks to improve image search engine performance? (4) How to enable more query modalities to help users express their search intentions? At the same time, the studies on low-level image representation also shifted to local invariant features, which have been successful in finding duplicate images and robust to partial occlusions.

Furthermore, the fast growth of social networks and mobile computing in recent years also brings in many new data, scenarios, and exciting problems to image search. For example, it is reported that, in 2011, Flickr surpassed 6 billion photos, whereas Facebook accumulated 60 billions photos.[1] It would be interesting to study how to effectively utilize image tags labeled by users to improve image search, annotation, and ranking. The popularity of mobile phones equipped with cameras further makes the scenario of query-by-image more desirable in various domains, such as landmark and book cover recognition. Other sensors such as GPS, compass, and gyro, which have become standard components of a smart phone, have also greatly advanced mobile image search by providing rich context.

1.4 Growth of Dataset Scales in the Past 20 Years

While the fundamental problem of semantic gap remains open, we have witnessed the great success of commercial systems and research progresses on image search. In particular, if we look at the scales of the databases used to build image search systems or to perform experimental evaluations in the past 20 years, we can clearly see an impressive advancement of image search.

In the early work of 1990s, a database usually consisted of several thousands of images. For example, the QBIC system used only 1,000 images [Faloutsos and Taubin 1993]. The Appearance PhotoBook was tested on 7,562 images of approximately 3,000 people [Pentland et al. 1994]. The Netra system utilized 2,500 Corel images [Ma and Manjunath 1997]. The VisualSEEk system was tested for 12,000 color images [Smith and Chang 1997]. The WebSEEk system was probably the largest system before 2000, which indexed 513,323 images and videos [Smith and Chang 1996]. However, for efficiency, content-based image search is only performed within the search result list returned for a textual query.

[1] http://thenextweb.com/socialmedia/2011/08/05/flickr-hits-6-billion-total-photos-but-facebook-does-that-every-2-months/.

In contrast to thousands of images typically used in 1990s, the dataset scale has been greatly increased to millions and even billions since 2000. For example, Wang et al. [2001] developed a system called SIMPLIcity which indexed 200,000 images. Quack et al. [2004] built a large-scale image retrieval system, Cortina, for 3 million Web images. Wang et al. [2006b] and Li et al. [2006] utilized 2.4 million Web photos to solve the image annotation problem by a search and mining process. To validate the effectiveness of the data-driven approach to image annotation and object recognition, Torralba et al. [2008a] collected 80 million images and demonstrated an impressive performance by employing a simple k-nearest neighbor search and voting strategy. To further leverage the power of large-scale data, Wang et al. [2010b] further enlarged the image dataset scale to 2 billion and tackled the duplicate cases for high accuracy image annotation.

We would like to emphasize that the dataset scale has a tremendous impact to image search. The impact is particularly significant after 2000. The explosive growth of Web multimedia data not only brings in many new technical challenges, but also provides a huge repository of training data to image search. In fact, many technical breakthroughs were caused by the availability of big data, which allows us to do brute force artificial intelligence (AI). Meanwhile, the user need of easy accessing, retrieving, and managing their interested personal and Web images becomes extraordinarily strong. As a result, the scale of the databases used in the research of image search has consistently increased from thousands in 1990s to millions and billions after 2000.

It is worth noting that though our focus is mainly on large-scale image search, there are still many open problems, for example, object retrieval and recognition, which require more in-depth studies on smaller-scale datasets and in return contribute to image search with more effective visual features.

1.5 Technical Breakthroughs around 2000

In 2000, Smeulders et al. [2000] published the survey paper "Content-based image retrieval at the end of the early years". Indeed, the year 2000 is a turning point for image search. We would like to highlight several technical breakthroughs around 2000.

—*Feature*. In 1999, Lowe [1999] proposed a new method called the Scale Invariant Feature Transform (SIFT) for local feature detection and description, which essentially addressed the problem of local patch matching. Its proven accuracy and efficiency in visual recognition and image retrieval have made a great impact on the research of low-level image representation, which has been apparently shifted from global features to local features since 2000.

—*Index*. In 1999, Gionis et al. [1999] developed a locality-sensitive hashing (LSH) technique, an efficient solution for the high-dimensional k-nearest neighbor search problem, which is essential to utilizing millions or billions of images in a search system. Since then, many variants of hashing algorithms have been proposed to further improve the search performance by seeking more effective projections or introducing supervised information.

—*System*. In 2001, Google launched its image search system for 250 million Web images, which is a successful milestone of image retrieval in commercial applications. Though the search engine is mainly text-based, its scalability of indexing hundreds of millions of images and its effective use of Web image surrounding texts have attracted researchers to study the problems of image annotation (the interaction between texts and images), efficient index, and novel image search scenarios.

The technical breakthroughs of effective features, efficient indices, and successful systems together mark a start of a new era for image search and greatly impact the research after 2000.

The authors have witnessed the change from thousands to billions in the past 20 years as one of the active first-hand participants. In this article, we will share our observations, suggestions, and predictions in image search from four aspects: system framework, feature extraction, large-scale indexing,

and big data potential. Since there already exist excellent surveys for image retrieval in the early years [Rui et al. 1999; Smeulders et al. 2000; Lew et al. 2006; Datta et al. 2008], we will especially emphasize the astonishing increase of dataset scales from thousands to billions in the past 20 years and the technology breakthroughs that have made this happen.

2. FRAMEWORKS

In this section, we will introduce the frameworks of image search, its major components, and the technical challenges. We will mainly focus on a generic framework for Web image search, but will also discuss frameworks used in early years to give readers a complete view. We will see that, because of different objectives in different stages, the frameworks exactly reflect those objectives.

Early work in *the text-based stage* (1970–1990) is mainly based on DBMS. The major focus was to build a pictorial data system on top of a database by employing simple image processing algorithms to extract lines and elementary shapes from pictorial data. The images or pictorial data are usually stored in the database as blob data, and the features are usually converted to alphanumeric type for the ease of manipulation and retrieval using database languages. Therefore the framework of pictorial data systems is fairly simple due to the use of database systems.

In *the content-based stage* (1990–2000), the major focus of the research was on the visual content analysis for image retrieval. As the number of images used for image retrieval in 1990s was typically of thousands or tens of thousands, the visual features can be simply loaded into memory, and a linear scan is usually sufficient to return the search result in (near) real time. Therefore the dependency on a database system is greatly removed. However, if there are millions of images, the problem becomes considerably challenging and efficient indexing technologies are highly needed.

In *the web-based stage* (2000–present), the explosive growth of Web images has made the demand of searching and browsing Web images imperative. Consequently, much research has been done on the problems of indexing Web images and utilizing Web images for image annotation and visual concept learning. Figure 1 shows a comprehensive framework for Web image search in this stage. Compared with frameworks in 1990s, the new framework has several additional modules, including Web image crawlers, multimodality features, and user query logs. The framework works for both text-based and visual-based image retrieval, depending on how the index is designed.

In the *offline* part, the system starts from crawling images from the Web and stores the crawled images into an image database. Though the function of a Web image crawler is fairly simple in logic, it is not a trivial task to design and implement a good crawler in a real system, due to many practical issues such as Web image selection, incremental updating, re-crawling frequency, super fresh image fetching, etc. [Olston and Najork 2010]. For each stored Web image, the system extracts its features. Depending on the target application, the features may include textual features extracted from the image's surrounding text, visual features extracted from the image's content, and query log features computed from users' previous clicks. For example, one simple yet effective query log feature can be constructed by associating queries with their frequently clicked resulting images to provide more accurate descriptions. In addition, associating more similar queries identified by log-based query clustering can further help address the term mismatching problem [Wen 2009]. After all the images have been processed, an index will be built by the system. For textual features, inverted index is widely used for its proven efficiency in web search. For visual features, the index structure varies greatly and still remains an active research problem, which will be discussed in Section 4.

In the *online* part, the system loads the index into memory, maybe partially with the design of a cache strategy, and starts to serve for image retrieval. Depending on the features and indices used in the system, a user can submit a text query or an image query, and the system returns a list of resulting images. To help the user formulate his or her query intention, an interactive query interface can be

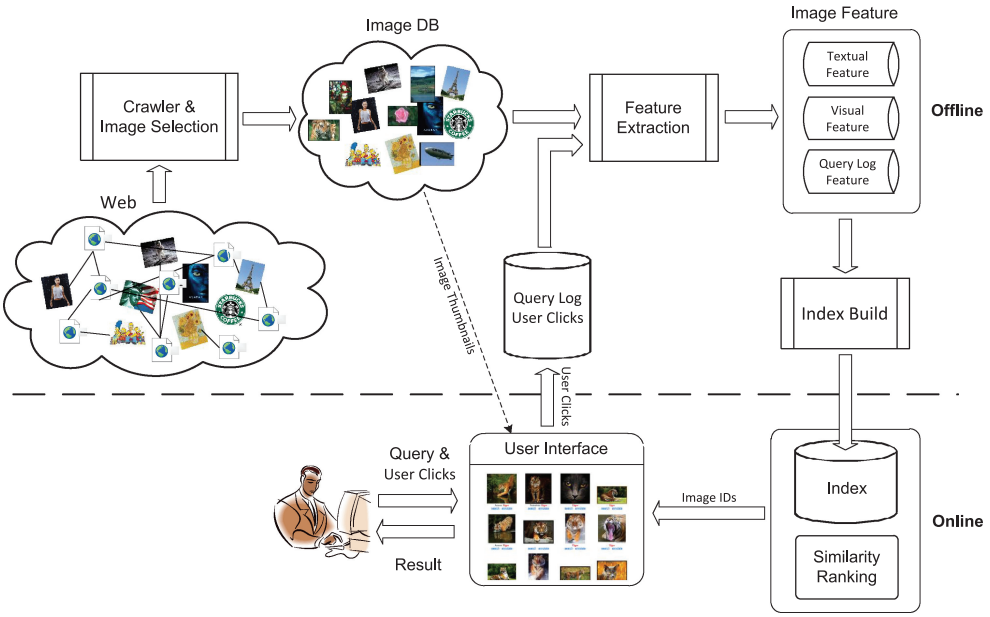

Fig. 1. A comprehensive image search framework in the 2000s.

developed, for example, query by color layout [Wang and Hua 2011], hand-drawn sketch [Cao et al. 2010b], or joint text and image suggestions [Zha et al. 2009]. When the user is browsing the result and is interested in an image, he or she may click the image thumbnail to see its full resolution version. The user clicks can be logged by the system and used to improve the search performance later. This is analogous to the explicit relevance feedback [Rui et al. 1998] but in an implicit way. For a commercial image search engine with numerous users, the user clicks are of great practical value to help the system improve the search relevance in many ways [Wen 2009].

Due to page limits, while this framework is of high level, it summarizes the key components in building different image search systems. For thousands of images in the content-based stage, the crawler and the query log are usually ignored, because of the static property of the data set and the lack of real users. All the remaining modules can be implemented efficiently using one server. For billions of images in the web-based stage, the framework is typically built as a parallel system consisting of tens or hundreds of servers to handle the scalability challenge. Moreover, both the features and the indexing techniques need to be deliberately designed to achieve the efficiency. We will have more discussions in the following sections.

3. FEATURES

A core research problem in image search is to compare two images and compute their similarities. The problem is generally not difficult for human beings, as our vision system can easily recognize structures and identify semantics from images, but it presents a grand challenge to computers. Although studies in neuroscience have shown that human brains also process external signals in a bottom-up manner [Trappenberg 2010], we still have very little knowledge how our vision system develops the capability of abstraction and generalization.

To mimic human vision systems, an image search engine should extract structural or syntax features, such as object boundaries and part relationships. Such kinds of features are closer to semantic

representation, but are not sufficiently robust to uncertainties in unseen images. In contrast, statistical features, such as the distribution of colors in an image, are more commonly adopted because of their robustness to structural changes including variances in translation, rotation, and scale.

Besides extracting visual features from images, there are also many ways to extract textual features by leveraging crowdsourcing. For example, one could parse surrounding texts as weak annotations of Web images, identify user-created tags from images in social media websites, engage users in games like ESP [Von Ahn and Dabbish 2004] and Peekaboom [Von Ahn et al. 2006] to let users collectively label images, and log user clicks in an image search engine to infer the association between a query word and its resulting images. Due to space limitations, we will not have a deeper discussion about extracting textual features, but mainly concentrate on visual feature extraction.

Because surveys for image retrieval in early years already exist [Rui et al. 1999; Smeulders et al. 2000; Datta et al. 2008], this section will devote to local features which have greatly changed the research on visual representation after the year 2000 and provided efficient indexing solutions for billions of images, and only briefly touch upon the latest advances in global features.

3.1 Local Features

Local features have been studied since the early 1980s [Moravec 1981], but only widely used in image search after 2000 since Lowe [1999] developed an efficient solution called SIFT for detecting scale-invariant local interest points and a robust descriptor for reliably matching local features. As, by design, the local features are distinctive and invariant to many types of geometric and photometric transformations, corresponding regions in two images will have similar (ideally identical) vector descriptors [Mikolajczyk et al. 2005], and therefore the correspondences can be easily established. Since there are multiple regions in an image, this method is particularly robust to partial occlusions.

3.1.1 *Local Feature Detection.* Local feature detection is a prerequisite step in obtaining local feature descriptions. It tries to detect repeatable local structures, for example, edges, corners, and blobs, in an image. These local structures are more distinctive and stable than other structures in smooth regions and are expected more useful for image matching and object recognition.

The study of local feature detection can be divided into three stages: (1) *Corner point localization*, studied by Moravec [1981] and Harris and Stephens [1988] to detect corner points for solving the stereo matching problem to estimate the depth information in robot vision; (2) *scale selection*, first studied by Witkin [1983] and Koenderink [1984] for the scale-space representation, and then by Lindeberg [1998] for identifying intrinsic scales of local image structures using *normalized scale-space derivatives*; (3) *efficient solution*, developed by Lowe [1999] to approximate *Laplacian-of-Gaussian* [Lindeberg 1994] using *difference-of-gaussian* for efficiency.

As shown by its name, the SIFT detector is scale and translation invariant but not affine invariant, and SIFT descriptor achieves a certain property of rotation invariance by normalizing the orientation histogram. Despite of its limitation in affine invariance, SIFT has shown to be robust in image matching across a substantial range of affine distortion, change in 3D viewpoint and illumination, and addition of noise. As a result, SIFT feature has been widely used in a large number of applications, for example, object categorization, image retrieval, robust matching, and robot localization.

To obtain affine invariant features, several detectors have been developed. These improvements include the Harris-Affine detector and the Hessian-Affine detector developed by Mikolajczyk and Schmid [2004], an edge-based region detector and an intensity-based region detector developed by Tuytelaars and Van Gool [2004], the maximally stable extremal region (MSER) detector developed by Matas et al. [2004], and an entropy-based region detector developed by Kadir et al. [2004]. There are also other improvements over the feature detection speed, such as SURF [Bay et al. 2006] and FAST [Rosten and

Drummond 2006; Rosten et al. 2010], or the repeatability of the feature detection result, such as Rank-SIFT [Li et al. 2011]. We will not introduce these detectors in detail. Interested readers can refer to published papers [Mikolajczyk et al. 2005; Mikolajczyk and Schmid 2005; Tuytelaars and Mikolajczyk 2008; Li and Allinson 2008] for comprehensive reviews and comparisons of these detectors.

3.1.2 *Local Feature Description.* To represent points and regions, a large number of different local descriptors have been developed. SIFT descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location. These samples are then accumulated into orientation histograms (8 bins) summarizing the content over 4×4 subregions, resulting in a 128-dimensional feature vector. An extension to SIFT is the gradient location and orientation histogram (GLOH), proposed by Mikolajczyk and Schmid [2005]. It computes the SIFT descriptor for a log-polar location grid and leads to a 272 bin histogram. Similar to SIFT and GLOH, Dalal and Triggs [2005] developed another local patch descriptor called HOG, standing for histogram of oriented gradients, which evaluates well-normalized local histograms of image gradient orientations in a dense grid of a local patch. HOG is proven quite effective on human body detection [Dalal and Triggs 2005] and also shows a good performance for similar image retrieval [Shrivastava et al. 2011].

There are also many other improvements, for example, for deriving a more compact descriptor, such as PCA-SIFT [Ke and Sukthankar 2004], to speed up the computation of descriptors, such as SURF [Bay et al. 2006], or to improve the discriminative power of descriptors [Hua et al. 2007]. Given the page limits, for interested readers, please refer to the original papers about the details.

Local feature descriptors are particularly effective in detecting near identical local structures, which can be treated as a low-level repeatable visual patterns. For this reason, SIFT works well for duplicate image retrieval. On the other hand, to detect larger local structures related to semantic meanings, more sophisticated algorithms need to be developed with the help of large-scale training data.

3.1.3 *Efficient Image Retrieval Based on Local Features.* A typical way of utilizing local features in image search is to detect local invariant features from an image and describe the image as a bag of local descriptors. The similarity between two images can then be measured by matching the descriptors between the two images. This process can be accelerated by quantizing continuous local descriptors to discrete visual words and represent an image as a bag of visual words [Sivic and Zisserman 2003]. The proven effective indexing and ranking schemes developed for Web search engines can then be utilized, making it possible to handle billions of images. This has become a common framework for large-scale local feature-based visual search and duplicate image retrieval, for example, landmark recognition [Zheng et al. 2009] and duplicate-based image search to annotation [Wang et al. 2010b].

The major limitation of bag-of-words representation is its lack of spatial information, which greatly limits its discriminative power and usually leads to inaccurate search results. To improve the accuracy, geometric verification methods (e.g., RANSAC [Fischler and Bolles 1981]) have been proposed to identify the true matches as a post processing step [Philbin et al. 2007]. However, such methods can be only applied to top-ranked images in an initial search result, because of their high computational complexity. To incorporate the spatial information more efficiently, visual phrase-based representation has been proposed to encode spatial information into feature groups [Chum et al. 2009; Wu et al. 2009; Cao et al. 2010a; Zhang et al. 2009a, 2010, 2011]. For example, Wu et al. [2009] suggested employing the MSER detector [Matas et al. 2004] to bundle SIFT features into local groups, Zhang et al. [2009a] proposed selecting descriptive visual phrases based on a large-scale training data set and model their spatial contexts [Zhang et al. 2010], and Zhang et al. [2011] proposed encoding spatial layout information of visual words to the index structure to obtain geometry-preserving visual phrases. Such approaches greatly improve the discriminative power of individual features.

3.2 Recent Advances on Global Features

Visual feature extraction was the major focus in the early years of image retrieval. Most features proposed in 1990s are global features. That is, such features are statistics about color and texture from a whole image. Since 2000, we have seen more features taking into account both statistical and structural information.

For example, *GIST* feature was proposed by Oliva and Torralba [2001] to summarize Gabor filter responses for scene image recognition, and can be treated as a holistic representation of texture layout for scenes images. In practice, many research works directly use the responses of multiscale oriented filters as features for similar image retrieval or object recognition [Torralba et al. 2003, 2008b; Hays and Efros 2007, 2008]. *Local binary patterns* (LBP) shows that complementary information of local spatial patterns and contrast plays an important role in texture discrimination and retrieval, and demonstrates a superior performance for face recognition [Ahonen et al. 2004].

Two more examples are *color map* [Wang and Hua 2011] and *edgel representation* [Cao et al. 2011], both of which consider the structural information, and meanwhile provide an efficient index solution. Color map is a representation of structural color features to enable users to express their search intentions by indicating how the colors are spatially distributed in the desired images. The representation is very compact (about 80 bytes for an image) and highly efficient, having been successfully applied to search billions of images.[2] To build a large-scale sketch-based image search system, Cao et al. [2011] proposed converting edge pixels in a shape image to visual words, and developed an index structure called *edgel index* to speed up the search and make real-time response possible in a million-level database. The MindFinder system, which indexes 2.1 millions images with 6.5GB memory cost on a common server, shows a promising research direction of shape-based image retrieval [Cao et al. 2011] and recognition [Sun et al. 2012].

4. INDEXING

Since 2000, the explosive growth of Web images has greatly reshaped the research of image retrieval. Given billions of images readily available on the Web, utilizing Web image datasets has become very important for various multimedia content analysis problems. However, in the early 2000s, most systems suffer from a scalability problem and cannot scale to millions or billions of images, because it is difficult to build an efficient index for high-dimensional image features [Li et al. 2007]. This challenge has motivated a substantial number of research attempts in large-scale image indexing.

According to the feature representations, high-dimensional image indexing techniques can be divided into two categories. The first category is for dense and continuous features, such as color histogram and local feature descriptors. The second category is for sparse bag-of-words features, which is widely used for local feature-based image retrieval. The difference between the two feature representations leads to different indexing techniques, which will be discussed in two separate sections.

4.1 Index for Dense Features

For dense features, the most widely used indexing techniques for conducting k-nearest neighbor search for image retrieval include kd-tree [Robinson 1981], and locality sensitive hashing (LSH) [Gionis et al. 1999; Datar et al. 2004]. The former is partition-based and the later is hash-based. Due to space limitations, we will not introduce kd-tree. Interested readers can refer to Weber et al. [1998] for a quantitative study of early indexing algorithms.

Locality sensitive hashing [Datar et al. 2004] is an *approximate nearest-neighbour* method of performing probabilistic dimension reduction of high-dimensional data. The key idea is to hash data points

[2]Microsoft Bing image search has enabled the feature of search by color since 2010.

using several hash functions constructed based on random projections to ensure that, for each function, the probability of collision is much higher for objects which are close to each other than for those which are far apart. Then, one can determine nearest neighbors by hashing the query point and retrieving elements stored in buckets containing that point.

The idea of hashing by random projection has received tremendous attentions in the multimedia and computer vision communities because it offers efficient solutions to billions of images in a search system. Following this idea, a great many variants of hashing algorithms have been proposed for further improving the search performance by seeking more effective projections or introducing supervised information. Torralba et al. [2008b] adopted learning approaches to convert the GIST descriptor to a compact binary code of a few hundred bits. Weiss et al. [2008] devised a new algorithm called spectral hashing to generate codes by thresholding a subset of eigenvectors of the Laplacian of the similarity graph. To optimize the search time, He et al. [2011] proposed a maximum entropy criterion and equivalently a minimum mutual information criterion to achieve the bucket balancing and thus optimize the search time. To utilize supervised information, Wang et al. [2010a] proposed a semisupervised hashing (SSH) framework that minimizes empirical error over the labeled set and an information theoretic regularizer over both labeled and unlabeled sets.

Being aware of the availability of the original feature of a given query in most retrieval applications, Dong et al. [2008] and Gordo and Perronnin [2011] proposed asymmetric distances to rank binary codes by taking advantage of the position information of the query point in the feature space. Zhang et al. [2012a] proposed a query-sensitive ranking method (QsRank) for PCA-based hash codes [Wang et al. 2006a], which not only considers the original query feature, but also models the statistical properties of the target ϵ-neighbors in the hash code space. Unlike Hamming distance, the asymmetric distances and the QsRank algorithm do not convert queries to binary hash codes when computing the ranking score, and thus the resulted ranking score is more effective than Hamming distance.

4.2 Index for Sparse Bag-of-Words Features

Bag-of-words representation was motivated by the success of Web search. The basic idea is to map image features to words [Sivic and Zisserman 2003; Philbin et al. 2007] and convert image retrieval problems to text retrieval problems, essentially making it possible to index billions of images using the proven effective indexing and ranking schemes developed for Web search.

To develop this type of image search engines for a large-scale database, there are still many technical challenges and problems that need to be addressed [Li et al. 2007]. The challenges include (1) how to construct a representative and discriminative visual *vocabulary*; (2) how to efficiently find relevant images for a *long query* of hundreds of visual words, in comparison to short queries in text search engine; (3) how to measure the *content quality* of Web images for efficient cache design and quality improvement; and (4) how to define a good *image similarity measure* by taking into account the spatial information of image words.

To address these technical challenges, much work has been done in recent years, especially since 2006. For example, to generate a larger and more discriminative vocabulary, Nister and Stewenius [2006] developed an efficient scheme to quantize local feature descriptors into a *hierarchical vocabulary tree* which leads to a dramatic improvement in retrieval quality. The most significant property of this scheme is that the tree directly defines the quantization. The quantization and the indexing are therefore fully integrated, essentially being one and the same.

Another important issue is the *long query* problem. Inverted index structure is particularly efficient for short queries (e.g., 3–5 terms), whereas an image query represented by bag-of-features usually contains hundreds of visual terms. This difference makes the inverted index structure inappropriate to index images. For example, given a query with three terms, all images containing them can be

obtained by intersecting three inverted lists. However, given a real image query which consists of 1,000 visual terms, we have to intersect 1,000 inverted lists and will very likely get an empty set. To address this problem, Zhang et al. [2009b] proposed decomposing a document-like representation of an image into two components, one for topic distribution and the other for residual information preservation. The topic distribution is a low-dimensional and dense feature, which can be indexed by technologies developed for dense features, and the residual information can be stored as metadata for ranking. The computing of similarity of two images can be transferred to measure similarities of their components. The good properties of the two components make both index and retrieval very efficient.

In comparison, the problem of *content quality* is less touched. The problem is related to image quality assessment, which is mainly based on the content analysis. Due to the lack of effective feature representation, it is extremely hard to develop a quality measure to match with the human perception about image quality. For Web images, we can resort to signals, such as pagerank, user click count, and website quality. But for generic images, quality assessment is still an open problem without an effective solution. There still remains a large improvement space for this problem.

4.3 Practical Index Solution for Heterogeneous Features

Practical image search systems are usually not limited to just visual features. Either Web images or mobile images are associated with rich contextual features, for example, surrounding text, user-clicks, location, social and mobile context, etc. How to effectively utilize such heterogeneous features and meanwhile build an efficient index presents a grand challenge to image search systems.

Although algorithms can be developed to project heterogeneous features to a common space and make them comparable, a more simple and practical solution can be adopted by utilizing a hybrid structure of an inverted index and a forward metadata index. Features that can be directly compared with query terms are usually put into inverted lists for fast candidate image selection, and other features are stored as metadata for providing more ranking features. For example, features such as image surrounding text, visual words, and content-based image categorization results can be put into inverted lists, whereas other features such as image resolution, picture taken time, location, and the number of faces can be stored in the metadata index.

Given a query, the system first looks up the inverted index to find inverted lists that match with the query terms, and then selects candidate images. After that, the system will fetch metadata, calculate a ranking score for each candidate image, and return the top-ranked images to users. To develop a ranking function that can incorporate all the features, a learn-to-rank algorithm [Liu 2009] can be utilized with the help of human labeling results.

Inspired by Web search engines, the architecture of inverted index, forward metadata index, and ranking function is generally highly scalable and preferred in practice.

5. BIG DATA, BIG POTENTIALS

As we have discussed in Section 1.4, the dataset scale has had a great impact on image search, particularly since 2000. While researchers continue working on the fundamental problem of visual representation for image content analysis and similarity measures, we have seen exciting progresses and new potentials for leveraging the Web as a huge repository for solving problems which were almost impossible to solve in the past. The strategy of utilizing massive data is analogous to brute force AI. As Web-scale image datasets with weak labels embody enormous amount of knowledge, an efficient computational model can empower computers to mimic human intelligence to a certain extent.

This section will highlight some recent work in three categories: (1) data-driven approaches to image annotation and visual recognition; (2) visual pattern mining from massive data; and (3) large-scale image knowledge base construction. A common characteristics of these works is that they all leverage

	2.4 M	80M	2B		2.4M	80M	2B
	(no results)	(no results)	*iranian actress,* older actresses, **nazanin boniadi** (4 dups)		(no results)	(no results)	**Stained Glass** (3 dups)
	(no results) (4 dups)	**iphone, apple,** Inhandhome, Lifestyle, Idevphone, *Devices* (20 dups)	*apps, ipod touch, standout macs iphone,* **apple iphone hand** (181 dups)		(no results) (3 dups)	**Seal** (5 dups)	*seal bashing, Beautiful Animals, endangered seals,* cutest seal in the world, **harper seal** (40 dups)

Fig. 2. Arista annotation examples vs. dataset size. Bold-faced tags are perfect terms labeled by human subjects and italic ones are relevant terms. The numbers of duplicates detected in each dataset are shown in parenthesis. For comparison, the number 2.4M was used in Wang et al. [2006b], 80M was used in Torralba et al. [2008a], and 2B was used in Wang et al. [2010b]. This figure suggests that a larger dataset size ensures more accurate tags.

large-scale Web images as well as their surrounding texts, and have to address two technical challenges: how to handle the scalability problem and how to deal with the noisy data.

5.1 Data-Driven Approaches to Image Annotation and Visual Recognition

In contrast to model-based approaches which mainly use mathematical models to learn and summarize visual concepts for image annotation and visual recognition, data-driven approaches are more dependent on large-scale data, and complement model-based approaches from a different angle by employing a k-nearest neighbor search strategy for image analysis. In this way, efficient image search is not the goal, but becomes the means to solving other problems.

5.1.1 *Search-Based Image Annotation.* An early study of search-based image annotation was conducted by Wang et al. [2006b] and Li et al. [2006]. In this work, the authors proposed reformulating image annotation as a two-step process: given an input image, first search for a group of similar images in a large-scale Web image database, and then mine key phrases extracted from the surrounding texts of the resulting images. In this work, 2.4 million Web images were crawled from several photo forum websites and a system was built to index the images as a knowledge base for image annotation.

To explore the potential of a much larger dataset, Torralba et al. [2008a] collected 80 million images from the Web for nonparametric object and scene recognition. Given an input image, the system searches in the 80 million image database to find its k-nearest neighbors and uses the associated labels to infer the semantic classes of the input image. The study confirmed the effect of larger database size for increasing the image classification accuracy.

Wang et al. [2010b] further advanced the state of the art by investigating the performance of image annotation on a dataset of two billion Web images. The study shows that for such a sufficiently large-scale data set, a near-duplicate search-based approach can generate accurate keywords for images which have duplicates in the database, and therefore is capable of addressing many difficult concepts, which cannot be effectively represented by existing visual features. Figure 2 shows a few examples of annotation results generated by three datasets of different sizes. The study also shows that about 8.1% Web images have more than ten duplicates. Considering the total number of two billion images, 8.1% corresponds to about 160 million images which can serve for many applications and researches, for example, building a celebrity dataset [Zhang et al. 2012b] or constructing a large-scale image knowledge base [Wang et al. 2012].

5.1.2 *Instance-Based Visual Recognition.* Instance-based recognition refers to recognizing a specific object instance by matching a query image with reference images in a large-scale database. Such an approach essentially utilizes the nearest neighbor search method and therefore highly depends on two factors: a large-scale database with good labels, and a reliable (and efficient) matching algorithm to compute the similarity of two images. Thanks to the Internet era, the availability of big data on the Web has greatly advanced the instance-based recognition, leading to successes in both research and commercial applications.

Fan et al. [2005] developed a Photo2Search system to support users to search for relevant information on the Web via photos of what they see, for example, a picture of a restaurant or a hotel. They collected several millions of images with latitude/longitude metadata and indexed the images using local features. Hays and Efros [2008] studied a similar problem of inferring the geo-location information of generic images by utilizing a dataset of over 6 million GPS-tagged images from the Internet. Zheng et al. [2009] developed a landmark recognition engine for 5,312 landmarks in 144 countries. In this work, the authors mined a comprehensive list of landmarks from 20 million GPS-tagged photos and an online tour guide webpages, and collected reference images from photo-sharing websites and image search engines. The recognition is performed by local feature matching of query image against reference images, based on the nearest neighbor principle. The experiments demonstrate that the engine can deliver a satisfactory recognition performance with high efficiency.

The long-term research on visual search and recognition has resulted in impressive commercial systems. For example, SnapTell[3] offers visual product search technologies for mobile phones to let users take a photo of the cover of any CD, DVD, book, or video game, and find ratings and pricing information online from Google, Amazon, and eBay. The company has a database of about 5 million+ products. In December 2009, Google launched Google Goggles[4], an image recognition application on mobile phones. Currently the system can identify landmarks [Zheng et al. 2009] and various labels, allowing users to learn about such items without needing a text-based search. In 2012, Kooaba relaunched Shortcut[5], which is literally a shortcut between real life and the Internet: take a picture of what you are reading in a newspaper or magazine and instantly get connected to the digital version.

5.2 Visual Pattern Mining from Massive Data

The success of local invariant feature detection in the past decade has greatly inspired the research on local structure-based feature representation, hoping to detect visual patterns that are more related to semantic meanings. The progress is particularly promising due to the help of large-scale training data.

For example, Tsai et al. [2011] proposed the visual synset representation by mining a dataset of 200 million Web images and 300 thousand labels. The authors first constructed a dictionary containing 300 thousand common English words which are frequently used in Web search. Then for each word, up to 1,000 images were downloaded from Google Image Search and further partitioned into visually similar groups. The images in the same group is formed as a visual synset, representing a prototypical visual concept. The visual synsets (with the trained SVM classifiers) can be treated as elementary semantics that are computationally detectable in unseen images. The authors demonstrated the superior performance of the visual synset approach for large-scale image annotation.

Another direction is to learn high-level and class-specific features from unlabeled massive images, typically using deep learning algorithms. A notable progress is reported in Le et al. [2012]. In this work, the authors utilized a nine-layered locally connected sparse autoencoder with pooling and local

[3]http://techcrunch.com/2009/06/16/image-recognition-startup-snaptell-acquired-by-amazon-subsidiary-a9com/.
[4]http://techcrunch.com/2009/12/07/google-goggles/.
[5]http://blog.kooaba.ch/2012/02/a-better-paperboy-introducing-kooaba-shortcut/.

Table I. Comparison between ImageNet and ImageKB

	#Concepts	#Images	Precision	Ontology Used	#Concepts in Ontology	Overlap with Query Log
ImageNet	21,841	14.2M	99.7%	WordNet	117,023	1.85%
ImageKB	0.52M	235.3M	80.0%	NeedleSeek	12.83M	5.16%

Concept: refer to synset in WordNet or item in NeedleSeek.
Precision: the numbers were evaluated on 80 random synsets in ImageNet, and 150 entities (just for top 10 images) in ImageKB.
Query Log: a six-month query log from Bing.
Overlap with Query Log: the intersection ratio between ontology and query log, in terms of exact match.

contrast normalization on a large dataset of 10 million images. The model has one billion connections and was trained on a cluster of 16,000 cores for three days. In their experiments, the authors obtained neurons that function as detectors for faces, human bodies, and cat faces. This work shows that it is possible to train neurons to be selective for high-level concepts without having to label images, making it easy to leverage as many as possible training images freely available on the Web.

5.3 Building Image Knowledge Base by Matching Images with an Ontology

An image knowledge base providing representative images for every visual entity in an immense ontology of human knowledge would be of great value for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing training and benchmarking data for such algorithms [Deng et al. 2009].

Motivated by this vision, researchers have started to create large image databases [Torralba et al. 2008a; Deng et al. 2009; Wang et al. 2012]. For example, the 80 million tiny image database was collected through querying search engines with items in a vocabulary of 75,062 nonabstract nouns selected from WordNet [Miller 1995]. Because the surrounding texts of Web images are in nature imprecise and noisy, images returned by search engines are usually irrelevant and inaccurate.

Deng et al. [2009] built a new database called ImageNet and resorted to crowdsourcing to clean irrelevant and inaccurate images. ImageNet is a large-scale ontology of images built upon the backbone of the WordNet structure, aiming to populate the majority of the 80,000 synsets of WordNet with an average of 500–1000 clean and full resolution images. The first stage of the construction of ImageNet involves collecting candidate images for each synset by querying several image search engines. In the second stage, every candidate image for a given synset is verified by multiple human labors recruited using the service of Amazon Mechanical Turk.[6] As of September 2012, ImageNet has processed 21,841 synsets and 14.2 million images.[7] Because it is much larger in scale and diversity and much more accurate than its predecessors, ImageNet has been widely used in many research works.

Wang et al. [2012] developed another image knowledge base called ImageKB, which is a graph representation of structured entities, categories, and representative images, as a new basis for practical image indexing and search. In contrast to ImageNet, ImageKB is automatically constructed via a both bottom-up and top-down, scalable approach that efficiently matches 2 billion Web images onto NeedleSeek,[8] an ontology with millions of entity nodes [Shi et al. 2010]. The approach consists of identifying duplicate image clusters from billions of images, discovering definitive text to represent an image cluster, and identifying representative images for an entity. A comparison between ImageNet and ImageKB is shown in Table I. As the construction of ImageKB does not depend on human labors, the process is much more efficient. However, the precision of ImageKB is apparently lower than that of ImageNet, which also shows the value of human labeling.

[6]http://aws.amazon.com/mturk/.
[7]http://www.image-net.org/.
[8]http://needleseek.msra.cn/.

6. CONCLUDING REMARKS AND FUTURE DIRECTIONS

From the reviews on image search in the past 20 years, we can see that the technical breakthroughs around the year 2000 have resulted in remarkable progresses in the past ten years, with image features shifting from global to local, indexing techniques evolving from partition-based to hash-based, and data sets scaling up from thousands to billions. Besides continually providing a huge repository of training data, the explosive growth of big data will bring greater technical challenges to image search and also help advance the research on effective visual representation, efficient index solution, heterogeneous data fusion, and user feedback utilization. At the end of 20 years of image search research, we would like to discuss a few future directions, hoping to see more promising progresses in the next ten years.

6.1 Effective Visual Representation

Effective visual representation is the foundation of image retrieval for both image content analysis and similarity measures. In the past 20 years, tremendous efforts have been made for finding better image representations. However, the current status is still unsatisfactory and the semantic gap problem remains a fundamental hindrance. This is because a machine vision system can only mimic our human vision system by processing pixel values of an image in a bottom-up manner, which is still far from a complete understanding of the mechanism of our human vision system for image interpretation. Any breakthroughs in this direction will need long-term and interdisciplinary researches in cognitive science, computer vision, and machine learning.

If we look back at the technical breakthroughs in the past ten years, local feature detection was a great success, which essentially solved the problem of duplicate structure detection. Motivated by this success, can we move a step further to develop more effective visual features to measure and detect similar or highly similar patterns? We have seen promising progresses on scene image representation [Oliva and Torralba 2001], visual pattern mining [Tsai et al. 2011], dense matching [Liu et al. 2011], and deep structure learning [Le et al. 2012]. The challenge is how to make them computationally more practical and efficient in visual representation. It would be exciting to see another visual representation breakthrough in the next ten years.

6.2 Image Knowledge Base Construction

In the communities of natural language processing, information retrieval, and knowledge discovery, ontologies have been heavily studied and developed for text understanding. Example ontologies are WordNet[9], Open Directory Project (ODP)[10], NeedleSeek[11], and FreeBase.[12]

Likewise, we have seen great progresses on building image knowledge bases in the past few years [Deng et al. 2009; Wang et al. 2012]. The problem is essentially how to connect billions of Web images (or even more) with a large-scale ontology of millions of entities, which requires a deep understanding of the semantics of images and an efficient matching algorithm. Deng et al. [2009] addressed this problem by first utilizing search and then resorting to crowdsourcing, whereas Wang et al. [2012] leveraged the redundancy of surrounding texts of duplicate images and developed an automatic approach to matching duplicate image clusters with an ontology.

Despite the exciting progresses, the current image knowledge bases are still relatively small, in the order of hundreds of millions. An ideal case would be to include all the images (or as many as possible) on the Web in an image knowledge base, and then images could be organized in a more structured way,

[9]http://wordnet.princeton.edu/.
[10]http://www.dmoz.org/.
[11]http://needleseek.msra.cn/.
[12]http://www.freebase.com/

differently from the flatten inverted index widely used in commercial image search engines. Towards this goal, an interesting problem would be how to grow the current image knowledge bases. Can we start from the seed images in current knowledge bases and propagate their semantics to more images? This will require reliable similarity measures and highly efficient algorithms.

6.3 Implicit User Feedback and Crowdsourcing

The commercial image search engines emerging from the year 2000 have greatly changed the research on relevance feedback. Those commercial systems do not expect users to provide explicit feedbacks, but are capable of collecting a huge number of user feedbacks in an implicit way by logging queries, user clicks, and even mouse hovers on resulting images. Treating user queries as a type of crowdsourcing and user clicks/mouse hovers as implicit feedbacks, the accumulated logs could be used to derive more effective image features for improving the relevance ranking.

However, despite their proven effectiveness, the user click data are in nature sparse and noisy, especially for tail queries. In this case, can we leverage the query logs of top queries to improve the tail queries? At a first glance, this problem looks similar to semisupervised learning, which typically employs a similarity graph and tries to propagate information from labeled nodes to unlabeled nodes via the graph. But for our problem, a grand challenge is how to handle the scalability issue. For billions of images, the cost of constructing a similarity graph will become computationally unaffordable. Therefore, approximate algorithms need to be studied for the purpose of efficiency. In addition, a reliable image similarity measure for constructing the similarity graph also needs to be thoroughly investigated. Considering the great value of user click data and query logs, any theoretical or engineering breakthroughs will lead to significant improvement of image search.

6.4 Mobile Image Search

An interesting challenge that content-based image retrieval (CBIR) usually has to face is why a user wants or needs to search for more similar images when he or she already has one image at hand. The scenario of query-by-image-example seems not particularly convincing to most normal users.

The emerging mobile image search may justify the query-by-image-example scenario, and also provides rich context information along with a query image. Queries on mobile devices with both content and context have greatly changed the traditional CBIR which purely depends on image content. For example, when a user submits a query image from his/her mobile phone to a search system, the system will possibly also know the query context such as location, time, motion speed, acceleration, direction, lighting condition, background noise, and touch input. In this case, how can a system leverage such context to infer a user's search intention and provide a better result? Much research work has been done in the literature to address this problem, but few has been proven convincing in real systems. Clearly more in-depth researches are needed. Given the increasing popularity of mobile phones, it is widely believed that mobile image search will become more important and pervasive. Google Goggles[13] and Google Glass[14] are two early examples showing the future of mobile image search.

6.5 Creative Multimedia Interface

Compared with text-based Web search, image search results are more visual and more attractive, and their meanings can be understood instantaneously by users. Taking into account such advantages, can we invent more creative interfaces to present image search results? Current image search engines have attempted to organize the search results into more informative structures [Jing et al. 2006, 2009,

[13]http://www.google.com/mobile/goggles.
[14]https://plus.google.com/+projectglass.

2012], or provide more intuitive interfaces to let users express their search intentions [Wang and Hua 2011]. We predict that with more visual features enabled in image search systems, more creative user interfaces will be developed in the near future.

This article has discussed the search problem for large-scale image datasets. Beyond search, there are also other creative ways to utilize large-scale image datasets, for example, browsing, navigation, and entertainment. A good example is Pinterest[15], a pinboard-style social photo sharing website, which develops a creative interface and platform to encourage users to create image collections, share interesting images, and browse other pinboards for inspiration. Creative multimedia interfaces not only require good interface designs, but also demand in-depth and multidisciplinary researches. Any breakthroughs in this aspect will be exciting and worth anticipation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Wei-Ying Ma for his visionary long-term support and encouragement, and Xin-Jing Wang, Changhu Wang, Xirong Li, and Zhiwei Li for their years of collaboration with the authors on the large-scale image search and annotation problems.

REFERENCES

AHONEN, T., HADID, A., AND PIETIKÄINEN, M. 2004. Face recognition with local binary patterns. In *Proceedings of the 8th European Conference on Computer Vision.* 469–481.

AOA. 2006. Good vision throughout life. http://www.aoa.org/x9419.xml.

BAY, H., TUYTELAARS, T., AND VAN GOOL, L. 2006. Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*. 404–417.

BLASER, A. 1979. Database techniques for pictorial applications. In *Database Techniques for Pictorial Applications*, Lecture Notes in Computer Science, vol. 81, Springer, Berlin.

CAO, Y., WANG, C., LI, Z., ZHANG, L., AND ZHANG, L. 2010a. Spatial-bag-of-features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3352–3359.

CAO, Y., WANG, C., ZHANG, L., AND ZHANG, L. 2011. Edgel index for large-scale sketch-based image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 761–768.

CAO, Y., WANG, H., WANG, C., LI, Z., ZHANG, L., AND ZHANG, L. 2010b. Mindfinder: Interactive sketch-based image search on millions of images (demo). In *Proceedings of the International Conference on Multimedia*. ACM, 1605–1608.

CHANG, N. AND FU, K. 1980a. Query-by-pictorial-example. *IEEE Trans. Softw. Eng.* 6, 519–524.

CHANG, N. AND FU, K. 1980b. A relational database system for images. In *Pictorial Information Systems*, Lecture Notes in Computer Science, vol. 80, Springer, Berlin, 288–321.

CHANG, S. AND KUNIL, T. 1981. Pictorial data-base systems. *Computer 14,* 11, 13–21.

CHANG, S., YAN, C., DIMITROFF, D., AND ARNDT, T. 1988. An intelligent image database system. *IEEE Trans. Softw. Eng. 14,* 5, 681–688.

CHUM, O., PERDOCH, M., AND MATAS, J. 2009. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 17–24.

DALAL, N. AND TRIGGS, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 1, IEEE, 886–893.

DATAR, M., IMMORLICA, N., INDYK, P., AND MIRROKNI, V. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the 20th Annual Symposium on Computational Geometry*. ACM, 253–262.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40,* 2, 5.

DENG, J., DONG, W., SOCHER, R., LI, L., LI, K., AND FEI-FEI, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

DONG, W., CHARIKAR, M., AND LI, K. 2008. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *Proceedings of the 31st Annual ACM SIGIR Conference*. ACM, 123–130.

[15]http://pinterest.com/.

FALOUTSOS, C. AND TAUBIN, G. 1993. The QBIC project: Querying images by content using color, texture, and shape. In *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases*. Vol. 1908, 173–187.

FAN, X., XIE, X., LI, Z., LI, M., AND MA, W. 2005. Photo-to-search: Using multimodal queries to search the Web from mobile devices. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM, 143–150.

FISCHLER, M. AND BOLLES, R. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24,* 6, 381–395.

GIONIS, A., INDYK, P., AND MOTWANI, R. 1999. Similarity search in high dimensions via hashing. In *Proceedings of the International Conference on Very Large Data Bases*. 518–529.

GORDO, A. AND PERRONNIN, F. 2011. Asymmetric distances for binary embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 729–736.

HARRIS, C. AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*. Vol. 15, 50.

HAYS, J. AND EFROS, A. 2007. Scene completion using millions of photographs. *ACM Trans. Graph. 26*.

HAYS, J. AND EFROS, A. 2008. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

HE, J., CHANG, S., RADHAKRISHNAN, R., AND BAUER, C. 2011. Compact hashing with joint optimization of search accuracy and time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 753–760.

HUA, G., BROWN, M., AND WINDER, S. 2007. Discriminant embedding for local image descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1–8.

JAIN, R. 1993. NSF workshop on visual information management systems. *SIGMOD Record 22,* 3, 57–75.

JING, F., WANG, C., YAO, Y., DENG, K., ZHANG, L., AND MA, W. 2006. Igroup: Web image search results clustering. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*. ACM, 377–384.

JING, Y., ROWLEY, H., ROSENBERG, C., WANG, J., AND COVELL, M. 2009. Visualizing Web Images via Google image swirl. In *Proceedings of the NIPS Workshop on Statistical Machine Learning for Visual Analytics*.

JING, Y., ROWLEY, H., WANG, J., TSAI, D., ROSENBERG, C., AND COVELL, M. 2012. Google image swirl: A large-scale content-based image visualization system. In *Proceedings of the International World Wide Web Conference*. ACM, 539–540.

KADIR, T., ZISSERMAN, A., AND BRADY, M. 2004. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision*. 228–241.

KE, Y. AND SUKTHANKAR, R. 2004. PCA-sift: A more distinctive representation for local image descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE, 506.

KOENDERINK, J. 1984. The structure of images. *Biol. Cybernet. 50,* 5, 363–370.

LE, Q., MONGA, R., DEVIN, M., CORRADO, G., CHEN, K., RANZATO, M., DEAN, J., AND NG, A. 2012. Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning*.

LEW, M., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2,* 1, 1–19.

LI, B., XIAO, R., LI, Z., CAI, R., LU, B., AND ZHANG, L. 2011. Rank-sift: Learning to rank repeatable local interest points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1737–1744.

LI, J. AND ALLINSON, N. 2008. A comprehensive review of current local features for computer vision. *Neurocomputing 71,* 10, 1771–1787.

LI, X., CHEN, L., ZHANG, L., LIN, F., AND MA, W. 2006. Image annotation by large-scale content-based image retrieval. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*. ACM, 607–610.

LI, Z., XIE, X., ZHANG, L., AND MA, W.-Y. 2007. Searching one billion Web images by content: Challenges and opportunities. In *Proceedings of the International Workshop Multimedia Content Analysis and Mining (MCAM)*. 33–36.

LINDEBERG, T. 1994. Scale-space theory: A basic tool for analyzing structures at different scales. *J. Appl. Stat. 21,* 1–2, 225–270.

LINDEBERG, T. 1998. Feature detection with automatic scale selection. *Int. J. Comput. Vision 30,* 2, 79–116.

LIU, C., YUEN, J., AND TORRALBA, A. 2011. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Machine Intell. 33,* 5, 978–994.

LIU, T.-Y. 2009. Learning to rank for information retrieval. *Found. Trends Info. Retriev. 3,* 3, 225–331.

LOWE, D. 1999. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2, IEEE, 1150–1157.

MA, W. AND MANJUNATH, B. 1997. Netra: A toolbox for navigating large image databases. In *Proceedings of the International Conference on Image Processing*. Vol. 1, IEEE, 568–571.

MATAS, J., CHUM, O., URBAN, M., AND PAJDLA, T. 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Comput. 22,* 10, 761–767.

MIKOLAJCZYK, K. AND SCHMID, C. 2004. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision 60,* 1, 63–86.

MIKOLAJCZYK, K. AND SCHMID, C. 2005. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Machine Intell. 27,* 10, 1615–1630.

MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T., AND GOOL, L. 2005. A comparison of affine region detectors. *Int. J. Comput. Vision 65,* 1, 43–72.

MILLER, G. 1995. Wordnet: A lexical database for english. *Commun. ACM 38,* 11, 39–41.

MORAVEC, H. 1981. Rover visual obstacle avoidance. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 785–790.

NISTER, D. AND STEWENIUS, H. 2006. Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE, 2161–2168.

OLIVA, A. AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision 42,* 3, 145–175.

OLSTON, C. AND NAJORK, M. 2010. Web crawling. *Found. Trends Inf. Retriev. 4,* 3, 175–246.

PENTLAND, A., PICARD, R., AND SCLAROFF, S. 1994. Photobook: Content-based manipulation of image databases. In *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases*. Vol. 2185, SPIE.

PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

QUACK, T., MÖNICH, U., THIELE, L., AND MANJUNATH, B. 2004. Cortina: A system for large-scale, content-based Web image retrieval. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*. ACM, 508–511.

ROBINSON, J. 1981. The KDB-tree: A search structure for large multidimensional dynamic indexes. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, 10–18.

ROSTEN, E. AND DRUMMOND, T. 2006. Machine learning for high-speed corner detection. In *Proceedings of the 9th European Conference on Computer Vision*. 430–443.

ROSTEN, E., PORTER, R., AND DRUMMOND, T. 2010. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Machine Intell. 32,* 1, 105–119.

RUI, Y., HUANG, T., AND CHANG, S. 1999. Image retrieval: Current techniques, promising directions, and open issues. *J. Visual Commun. image Represen. 10,* 1, 39–62.

RUI, Y., HUANG, T., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol. 8,* 5, 644–655.

SHI, S., ZHANG, H., YUAN, X., AND WEN, J. 2010. Corpus-based semantic class mining: Distributional vs. pattern-based approaches. In *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL, 993–1001.

SHRIVASTAVA, A., MALISIEWICZ, T., GUPTA, A., AND EFROS, A. 2011. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph. 30,* 6.

SIVIC, J. AND ZISSERMAN, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1470–1477.

SMEULDERS, A., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Machine Intell. 22,* 12, 1349–1380.

SMITH, J. AND CHANG, S. 1996. Searching for images and videos on the World-Wide Web. *IEEE Multimedia Mag.*

SMITH, J. AND CHANG, S. 1997. Visualseek: A fully automated content-based image query system. In *Proceedings of the 4th ACM International Conference on Multimedia*. ACM, 87–98.

SUN, Z., WANG, C., ZHANG, L., AND ZHANG, L. 2012. Query-adaptive shape topic mining for hand-drawn sketch recognition. In *Proceedings of the 20th ACM International Conference on Multimedia*. ACM.

TORRALBA, A., FERGUS, R., AND FREEMAN, W. 2008a. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Machine Intell. 30,* 11, 1958–1970.

TORRALBA, A., FERGUS, R., AND WEISS, Y. 2008b. Small codes and large image databases for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

TORRALBA, A., MURPHY, K., FREEMAN, W., AND RUBIN, M. 2003. Context-based vision system for place and object recognition. In *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 273–280.

TRAPPENBERG, T. P. 2010. *Fundamentals of Computational Neuroscience*. Oxford University Press.

TSAI, D., JING, Y., LIU, Y., ROWLEY, H., IOFFE, S., AND REHG, J. 2011. Large-scale image annotation using visual synset. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 611–618.

TUYTELAARS, T. AND MIKOLAJCZYK, K. 2008. Local invariant feature detectors: A survey. *Found. Trends Comput. Graphics Vision 3,* 3, 177–280.

TUYTELAARS, T. AND VAN GOOL, L. 2004. Matching widely separated views based on affine invariant regions. *Int. J. Comput. Vision 59,* 1, 61–85.

VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319–326.

VON AHN, L., LIU, R., AND BLUM, M. 2006. Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 55–64.

WANG, B., LI, Z., LI, M., AND MA, W. 2006a. Large-scale duplicate detection for Web image search. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 353–356.

WANG, J. AND HUA, X. 2011. Interactive image search by color map. *ACM Trans. Intell. Syst. Technol. 3,* 1.

WANG, J., KUMAR, S., AND CHANG, S. 2010a. Semi-supervised hashing for large scale search. *IEEE Trans. Pattern Anal. Machine Intell. 6,* 1, 1.

WANG, J., LI, J., AND WIEDERHOLD, G. 2001. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Machine Intell. 23,* 9, 947–963.

WANG, X., ZHANG, L., JING, F., AND MA, W. 2006b. Annosearch: Image auto-annotation by search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2, IEEE, 1483–1490.

WANG, X., ZHANG, L., LIU, M., LI, Y., AND MA, W. 2010b. Arista-image search to annotation on billions of Web photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2987–2994.

WANG, X.-J., XU, Z., ZHANG, L., LIU, C., AND RUI, Y. 2012. Towards indexing representative images on the web. In *Proceedings of the 20th ACM International Conference on Multimedia*. ACM.

WEBER, R., SCHEK, H., AND BLOTT, S. 1998. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the International Conference on Very Large Data Bases*. 194–205.

WEISS, Y., TORRALBA, A., AND FERGUS, R. 2008. Spectral hashing. In *Proceedings of the Conference on Neural Information Processing Systems*.

WEN, J.-R. 2009. *Encyclopedia of Data Warehousing and Mining* 2nd Ed. IGI Global, Chapter Enhancing Web Search through Query Log Mining, 758–763.

WITKIN, A. 1983. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence.* Vol. 2, Morgan Kaufmann Publishers Inc., 1019–1022.

WU, Z., KE, Q., ISARD, M., AND SUN, J. 2009. Bundling features for large scale partial-duplicate Web image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 25–32.

ZHA, Z., YANG, L., MEI, T., WANG, M., AND WANG, Z. 2009. Visual query suggestion. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 15–24.

ZHANG, S., HUANG, Q., HUA, G., JIANG, S., GAO, W., AND TIAN, Q. 2010. Building contextual visual vocabulary for large-scale image applications. In *Proceedings of the International Conference on Multimedia*. ACM, 501–510.

ZHANG, S., TIAN, Q., HUA, G., HUANG, Q., AND LI, S. 2009a. Descriptive visual words and visual phrases for image applications. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 75–84.

ZHANG, X., LI, Z., ZHANG, L., MA, W., AND SHUM, H. 2009b. Efficient indexing for large scale visual search. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1103–1110.

ZHANG, X., ZHANG, L., AND SHUM, H. 2012a. Qsrank: Query-sensitive hash code ranking for efficient-neighbor search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2058–2065.

ZHANG, X., ZHANG, L., WANG, X.-J., AND SHUM, H.-Y. 2012b. Finding celebrities in billions of Web images. *IEEE Trans. Multimedia 14,* 4, 995–1007.

ZHANG, Y., JIA, Z., AND CHEN, T. 2011. Image retrieval with geometry-preserving visual phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 809–816.

ZHENG, Y., ZHAO, M., SONG, Y., ADAM, H., BUDDEMEIER, U., BISSACCO, A., BRUCHER, F., CHUA, T., AND NEVEN, H. 2009. Tour the world: Building a Web-scale landmark recognition engine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1085–1092.

ZHOU, X. AND HUANG, T. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst. 8,* 6, 536–544.

Received September 2012; revised March 2013; accepted March 2013