

Searching One Billion Web Images by Content: Challenges and Opportunities

Zhiwei Li, Xing Xie, Lei Zhang, and Wei-Ying Ma

Web Search and Data Mining Group
Microsoft Research Asia
{zli, xingx, leizhang, wyma}@microsoft.com

Abstract. Although content-based image retrieval has been studied for decades, most commercial image search engines are still text-based. However, there is a growing demand for techniques to support content-based image search at Web scale. In this paper, we propose an ambitious goal to searching one billion Web images by content, and discuss the major challenges and opportunities. We also present several important applications that can be greatly benefited by techniques to enable Web-scale image search by content. These applications include image copyright infringement detection, street-side photo search, and search-based image annotation. We believe that the insights presented in the paper are enlightening to researchers in this field, and any breakthrough we make in this space will lead to many impactful applications in the future.

Keywords: Content based image retrieval, web search, multimedia mining.

1 Introduction

Content-based image retrieval (CBIR) has been studied for decades. Although researchers have obtained a lot of promising results, the use of CBIR technologies in real commercial system is still very limited. On the other hand, the number of images on the Web keeps growing rapidly. It is reported that there are more than one million new images uploaded to Flickr [1] every day, and most commercial search engines have indexed over several billions of Web images based on the surrounding text. It seems that keyword-based image search engine is already sufficient for serving most people's information need. However, we believe that if billions of online images can be well utilized, it is possible to develop a more powerful search engine and lead to new and promising applications.

Most existing CBIR systems suffer from scalability problem and cannot scale to billions of photos. It is difficult to build effective index for high dimensional image features. Motivated by the success of web search engines, many researchers have tried to map image retrieval problems to text retrieval problems, hoping that the proven effective indexing and ranking schemes can be used to handle the scale. The basic idea is to map image features to words. Typically images are first represented by local features, and then by clustering, each local feature is mapped to a discrete keyword. Such image representation is called "bag of features," similar to "bag of words" for

document representation. With this representation, comparing two images become matching words in them, and therefore, a text-based search engine can be utilized to reduce the computational and memory cost.

To develop this type of image search engines, there are still many technical challenges and problems that we need to address:

- **Vocabulary:** what kinds of image features should be used? How to map them to words? The most generally utilized method is clustering. Some researchers also adopted a hierarchical clustering method to generate a vocabulary tree. But it is clear that we need to develop some kinds of visual language models to solve the problem.
- **Long query:** The reason why text search engine is effective is because text queries usually only contain a few words. So, the query-document matching can be conducted efficiently by inverted index. Although images can be represented by “bag of features,” the retrieval problem is still very different from text retrieval because query-by-example is actually equivalent to using a whole document as a query. So, the search is more like document-to-document matching. How can we deal with this kind of “long query” effectively?
- **Content quality:** Web search engine is effective because it can use link analysis to obtain quality and importance measurement (e.g. PageRank) for Web pages. For images, it is hard to obtain similar kind of measurement because the links are typically not directly associated with images. Without PageRank for images, we won’t be able to take advantage of many top-k search techniques typically used in web search, and it also will lead to the lack of efficient cache of index.
- **Relevance ranking:** The similarity measure between two images is quite different from text. How image words are weighted in computing the relevance. And how to deal with “word proximity” in images?
- **Distributed computing for Web-scale multimedia analysis:** Because of the large volume of image data we need to process and index, the system has to be a distributed system, consisting of hundreds of powerful servers. It is inevitably to confront with the challenges as in text-based search engines, such as fault tolerance, data redundant backup, auto configuration, etc.

To be able to response to a query within one second, the system has to employ a very efficient indexing solution, which is probably similar to inverted lists used in text-based search engines. The use of the indexing solution will make many algorithms depending on sequential scan of the whole database impractical. Therefore, most existing CBIR algorithms need to be re-evaluated on a Web-scale content-based image search system.

Due to the restrictions imposed by the indexing solution, the user interface for relevance feedback needs to be restudied. For example, it is not trivial to refine the search result by query point movement or distance function modification based on relevance feedbacks in an inverted index-like system. How to leverage users’ feedback either explicitly or implicitly will be an interesting research problem in a Web-scale content-based image search system.

It is not a trivial task to extract low level features from every image in a database containing one billion images. We need a flexible platform and infrastructure to provide large-scale data management and data processing capabilities. The infrastructure should facilitate the extraction and experimentation of various features and similarity measures for image search, so that it can help researchers and engineers to find a best practical solution by carefully evaluate the capabilities and limitations of different features and algorithms.

2 Applications

There are many important applications that can be benefited by the technology if we can effectively scale it up to search one billion image by content. In the following, we introduce a few related projects at Microsoft Research Asia.

2.1 Copyright Infringement Detection

Images are also one kind of intelligence property (IP) of their creators. However, with more and more user-created content on the Web, people may utilize other people's copyright images and videos without authorization. Thus, there is a need to develop a technique to help image owners to detect whether their images are used on the Web without their authorization or permission. Given a suspicious image, our task is to find the most similar image (or near-duplicate image) using our image search engine that indexes billions of Web images. Note that pirated images may be edited slightly, and the amount of images we have to match is very huge. Detecting such copyright infringement in a large image data collection is a non-trivial task.

2.2 Street-Side Photo Search

Mobile phones with embedded cameras are becoming popular nowadays and have huge growth potentials. Most current services for information acquisition on mobile devices are using text-based inputs. Nevertheless, sometimes it is difficult for users to describe their information needs in words. Instead of current flat query modes, camera phones can support much richer queries, not only text but also images. Therefore, it is important to develop a mobile search service that allows users to search for relevant information on the Web via the pictures taken on a mobile phone.

2.3 Search-Based Image Annotation

Although it has been studied for several years by computer vision and machine learning communities, image annotation is still far from practical. One reason is that it is still unclear how to model the semantic concepts effectively and efficiently. The other reason is the lack of training data, and hence the semantic gap cannot be effectively bridged.

With the growing number of images on the Web, it is possible to use the search technology to annotate unlabeled images. We can reformulate the image annotation problem as a novel two-step fashion: searching for semantically and visually similar images on the Web, and mining key phrases extracted from the descriptions of the

images [2],[3]. Intuitively, if a well annotated and unlimited-scale image database is available, then for any query image, we can find its near-duplicate in this database and simply propagate its annotation to the query image. In a more realistic case that the image database is of limited scale, we can still find a group of very similar images in terms of either global features or local features, extract salient phrases from their descriptions, and select the most salient ones to annotate the query image.

3 Summary

Searching one billion images by content is an ambitious and challenging direction. In this position paper, we have analyzed the major difficulties in indexing and searching a Web-scale image database. The challenges include how to construct a representative visual vocabulary, how to efficiently find relevant images for a long query, and how to compute “PageRank” for images for cache design and quality improvement. These challenges also motivated us to develop a flexible platform and infrastructure for “scale” experiment with various algorithms [4]. We also present several important applications related to Web-scale image search by content. We believe that the insights in this paper are enlightening to researchers in this field, and the large scale content-based image search will lead to many impactful applications in the future.

References

1. http://blog.flickr.com/flickrblog/2006/08/geotagging_one_.html Geotagging - one day later (August 29, 2006)
2. Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.Y.: Image Annotation by Large-Scale Content-based Image Retrieval, In: Proc. of ACM Int. Conf. on Multimedia, Santa Barbara, USA (2006)
3. Wang, X., Zhang, L., Jing, F., Ma, W.Y.: AnnoSearch: Image Auto-Annotation by Search. In: Proc of the International Conference on Computer Vision and Pattern Recognition (CVPR), New York (2006)
4. Wen, J.R., Ma, W.Y.: WebStudio - Building Infrastructure for Web Data Management, ACM SIGMOD (2007)