Signal Processing 124 (2016) 13–26

Contents lists available at ScienceDirect

Signal Processing

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

On very large scale test collection for landmark image search benchmarking

Zhiyong Cheng, Jialie Shen *

School of Information Systems, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 2 July 2015
Received in revised form
29 October 2015
Accepted 30 October 2015
Available online 20 November 2015

Keywords:
Large scale landmark image search
Performance evaluation

ABSTRACT

High quality test collections have been becoming more and more important for the technological advancement in geo-referenced image retrieval and analytics. In this paper, we present a large scale test collection to support robust performance evaluation of landmark image search and corresponding construction methodology. Using the approach, we develop a very large scale test collection consisting of three key components: (1) 355,141 images of 128 landmarks in five cities across three continents crawled from Flickr; (2) different kinds of textual features for each image, including surrounding text (e.g. tags), contextual data (e.g. geo-location and upload time), and metadata (e.g. uploader and EXIF); and (3) six types of low-level visual features. In order to support robust and effective performance assessment, a series of baseline experimental studies have been conducted on the search performance over both textual and visual queries. The results demonstrate importance and effectiveness of the test collection.

1. Introduction

In general, landmark refers to notable buildings (i.e. Statue of Liberty), architecture with special structure or meaning or purpose (i.e. Beijing National Stadium – "Bird's Nest") and famous scenic spots (e.g. Marina Bay in Singapore). Fig. 1 illustrates a few examples. Due to the attractive physical features or/and historical significance, landmarks frequently attract a lot of visitors, who are keen on taking the photos and share them with friends or/and family members via online social communities. Consequently, volume of landmark images increases tremendously in recent years and has accounted for a significant portion of online social images. In recent years, many different algorithms or systems have been developed to support automatic retrieval or visualization of landmark images [1–6]. In particular, large scale landmark image search emerge as important technical foundation for various real applications [7]. Consequently, numerous efforts have been devoted to improve the corresponding search systems' performance from different perspectives (e.g. retrieval effectiveness [8–12], visual classification [13], system performance evaluation [14], and result diversification [15,16]).

The technology advancement in landmark image search is largely dependent on studying and analyzing system performance. However, very limited work has been carried out on benchmarking dataset development for the purpose of comparing and evaluating relative algorithms and systems comprehensively. While the importance of the issue has been recognized in the multimedia retrieval and other related communities (e.g. computer vision and signal processing) and a few test collections have been published recently, they generally suffer from one or multiple weaknesses as follows: small scale, unclear definition about search task, lack of diversified landmarks views and

* Corresponding author.

http://dx.doi.org/10.1016/j.sigpro.2015.10.037
0165-1684/

Fig. 1. Examples of landmark images.

limited availability. The issues could be particularly severe when the researchers try to do robust cross-method comparisons. Due to the lack of quality collections, a popular solution taken by scholars is to construct their own datasets by leveraging online public resources, such as Flick[1] and Google image [8,15–20]. This can easily lead to very expensive and tedious dataset development process. More importantly, the use of self-constructed datasets makes it hard for other scholars to repeat the experimental studies and compare different methods to assess (1) the precise impacts of various systems and (2) identify the state-of-the-art.

In principle, the standard procedure for the performance evaluation of landmark image retrieval systems can include five basic steps: (1) construct a test collection; (2) define specific search tasks; (3) select search queries (text or/and visual queries) and generate associated ground truth; (4) run each test query through a particular landmark search system; and (5) assess the performance of the system via an empirical distribution of particular measurement metric (e.g. precision, recall and MAP ratio). All five steps are critical for the quality of performance evaluation. In this paper, our main focus is on how to develop very large scale of test collection. To achieve reliable, robust and effective system performance assessment, test collection construction needs to satisfy three key guidelines:

- Given that the size of image collections in many real photo sharing Websites have scaled to billions over the last few years, test collection's scale is required to be sufficiently big to generate statistical meaningful results.
- Real geographic locations in different countries and regions might have very diverse visual appearance and thus test collection should own comprehensive visual coverage of different geographic locations.

[1] https://www.flickr.com/