Multimed Tools Appl (2011) 51:77–98
DOI 10.1007/s11042-010-0630-z

Research and applications on georeferenced multimedia: a survey

Yan-Tao Zheng · Zheng-Jun Zha · Tat-Seng Chua

Published online: 5 November 2010

Abstract In recent years, the emergence of georeferenced media, like geotagged photos, on the Internet has opened up a new world of possibilities for geographic related research and applications. Despite of its short history, georeferenced media has been attracting attentions from several major research communities of Computer Vision, Multimedia, Digital Libraries and KDD. This paper provides a comprehensive survey on recent research and applications on online georeferenced media. Specifically, the survey focuses on four aspects: (1) organizing and browsing georeferenced media resources, (2) mining semantic/social knowledge from georeferenced media, (3) learning landmarks in the world, and (4) estimating geographic location of a photo. Furthermore, based on the current technical achievements, open research issues and challenges are identified, and directions that can lead to compelling applications are suggested.

Keywords Georeferenced media · Geotagged photo · Survey

Y.-T. Zheng (✉)
Institute for Infocomm Research, Singapore, Singapore
e-mail: yzheng@i2r.a-star.edu.sg, yantaozheng@gmail.com

Z.-J. Zha · T.-S. Chua
Department of Computer Science, National University of Singapore,
Singapore, Singapore

Z.-J. Zha
e-mail: zhazj@comp.nus.edu.sg

T.-S. Chua
e-mail: dcscts@comp.nus.edu.sg

1 Introduction

Recent years have witnessed the phenomenal advances of media-sharing services on the Internet, such as Flickr™ and Youtube™. Together with geotagging[1] facilities, these media repositories host sheer volume of georeferenced and community-contributed media resources, including documents, photos and videos, etc. For example, Flickr[2] hosts over 40 millions public georeferenced photos, while Wikipedia[3] lodges over 1 million geotagged articles [43, 44]. Collocated with temporal references and textual meta-data, these enriched multimedia have provided a unprecedented wealth of data to solve geographic-related multimedia and vision tasks that were unattainable in the past.

As shown in Fig. 1, in general, there exist three types of media with time- and georeferences on the Internet: (1) geotagged photos on photo-sharing websites like Flickr™, (2) georeferenced videos on websites like Youtube,[4] and (3) georeferenced web documents, like articles in Wikipedia and blogs in MySpace.[5] In the era of Web 2.0, the various georeferenced media are mostly socially generated, collaboratively authored and community-contributed. The time- and geo-references, together with text meta-data, reflect where and when the media was collected or authored, or the locations and times described by the media content. The enriched online multimedia resources open up a new world of opportunities to discover geographic related knowledge and information of our human society. For example, billions of geo- and time- referenced photos on Flickr[6] connect geography, time and visual information together and provide possibilities to discover visual patterns and knowledge of a particular geo-location.

Though mining on georeference multimedia is a recently emerging research topic, geographic data mining has been an active research field in the research community of knowledge discovery from databases (KDD) [5, 11, 29, 45, 61, 83]. In the domain of KDD, geographic data mining, or geographic knowledge discovery (GKD), refers to the process of extracting implicit knowledge, geospatial relations, rules and knowledge from massive georeferenced databases [13, 23, 29]. Since 1990s, digital geospatial data have gone through immense explosion, fueled by the technological developments of digital mapping, remote sensing and Global Positioning System (GPS), etc. Facing the sheer volume of digital geographical data, researchers from the KDD community developed various approaches for efficient geographical analysis and spatial relationship modeling [29, 45].

Compared to geographic data mining in KDD, recent research on online georeferenced multimedia differs in two-fold. First, geographic data mining usually

[1] Geotagging, or georeferencing, here refers to associating a media resource with geographical/location information.

[2] http://flickr.com

[3] http://wikipeida.org

[4] http://youtube.com

[5] http://myspace.com

[6] http://flickr.com

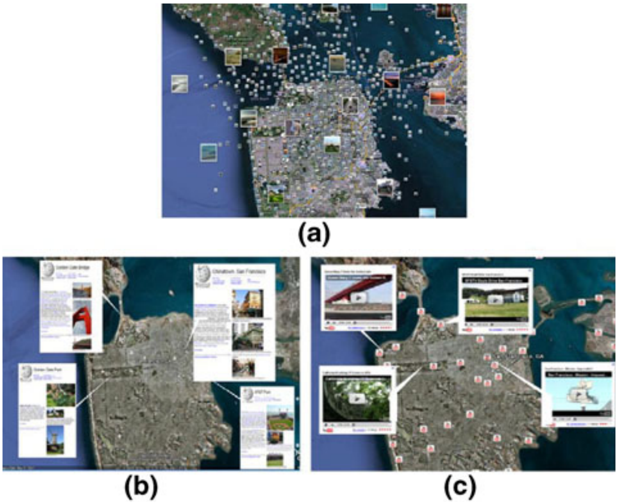

Fig. 1 Georeferenced multimedia documents in San Francisco area. **a** Georeferenced photos, **b** georeferenced wikipedia documents, and **c** georeferenced Youtube videos

works on specialized georeferenced databases, such as digital map of land usage, demographical database of social networks, etc. In contrast, the online georeferenced multimedia are community-contributed data that describe places, events, activities, and various aspects of people's life in a general manner. Second, geographic objects and spatial relationships in KDD studies are interpreted in spatio-temporal representations, while the online multimedia data present multi-modal information from visual, textual, geospatial and temporal channels. Due to the two differences above, existing geographic data mining methods can not be applied to georeferenced multimedia "as is".

The multi-modality and heterogeneity of online georeferenced multimedia have encompassed challenges not seen in traditional geographic data ming and attracted attentions of researchers from various community of KDD, Multimedia, Digital library and Computer Vision. This motivates us to survey the recent research on online georeferenced multimedia, and appraise what have been achieved and what are the challenges and directions that can lead to compelling applications. Specifically, we review the recent literature work in the following four aspects: (1) organizing and browsing georeferenced media resources, (2) mining semantic/social knowledge from georeferenced media, (3) learning landmarks in the world, and (4) estimating geographic location of a photo. The rest of the paper is organized as follows. Section 2 gives an overview of various methods of geotagging media resources, ranging from integrated hardware to software solutions. Section 3 surveys the recent research and applications on georeferenced multimedia resources on the Internet. In Section 4, we discuss the challenges and research directions. Finally, Section 5 gives conclusive remarks.

2 Collective geotagging

The voluminous georeferenced media on the Internet are a result of collective geotagging by the web community. Geotagging refers to the process of adding

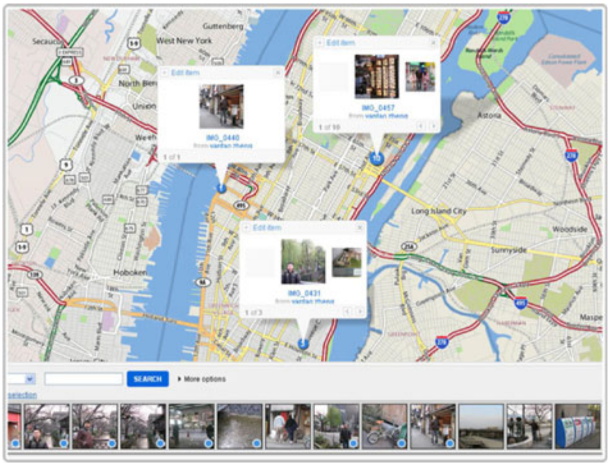

Fig. 2 The geotagging interface in Flickr. Users drag and drop photos to a location on the map to geotag them

geographical identification metadata to media resources, such as photographs, video, articles, websites and so on [39]. The metadata usually consists of GPS latitude and longitude coordinates, and sometimes, altitude, camera heading direction and place names [14, 85]. In general, the means of geotagging can be classified into two types: integrated hardware (automatic), and purely software solutions (manual) [17].

To date, an ever increasing amount of cameras and smart phones have been embracing GPS technologies, as the hardware cost becomes more trivial. These capturing devices save the GPS coordinates of the photo shooting location into the EXIF fields, together with other attributes like shutter times, focal lengths, etc. Besides GPS devices embedded in camera, GPS logger, a standalone GPS add-on for cameras, is an alternative means to log location information into photos. Being attached to a camera, a GPS logger can embed location information with photos through timestamp correlation and interpolation. In addition to GPS technologies, Wi-Fi, cellular radio, like GSM and CDMA, and other sensor networks have been used as hybrid positioning systems to determine the geo-location via triangulation [66, 77]. GPS and other geo-location acquisition hardware provide an automatic solution for geotagging photos and videos. However, till now, only a small portion of georeferenced media are geotagged via this means, as GPS-equipped cameras are far from prevalent.

Most georeferenced media on the Internet are retro tagged by web users manually via a geotagging software platform. To facilitate easy geotagging, commercial media-sharing services, including Picasa™, Zoomr,[7] Flickr, Panoramio, and Youtube, etc, have adopted map based tagging tools. In general, these geotagging tools allow a user to drag and drop photos to a location on the map, as shown in Fig. 2. The intuitive map and user-friendly interface render the geotagging a simple and straightforward process. However, the major limitation of such geotagging processes is that the accuracy in the location specification is not fine enough to identify the precise point where a photo has been shot. Currently, there exist no industry standards on tagging

[7] http://zoomr.com

and storing geotags of media. Most commercial media repositories store geotags in tag-based systems, similar to how text tags are stored.

Though the software platform requires manual intervention in geo-tagging photos, its simplicity has made it particularly popular and widely used in photo repositories like Flickr and Panoramio. In August 2006, geo-tagging facilities at Flickr started to operate; and by the year of 2007, more than 20 million geo-tagged photos have been uploaded at Flickr [44]. To date, Flickr receives more than 3 million geo-tagged photos per month. Similarly, Panoramio started in October 2005; and by March 2007, it has received more than 1 million geo-tagged photos [67].

3 Research and applications on georeferenced media

In this section, we survey the research work on georeferenced media from the following aspects: (1) organizing and browsing georeferenced media resources, (2) mining semantic/social knowledge from georeferenced media, (3) learning landmarks in the world, and (4) estimating geographical location of a photo.

3.1 Browsing, organizing and summarizing photo collections

Geographic location is one of the most important memory cues to recall people's past events [92]. The cognitive values of geo-location makes it extremely helpful in organizing, browsing and visualizing media collections, ranging from a single user's personal photo album to a global collection of digital media resources [64, 82]. In particular, Digital Libraries and Multimedia communities have been active in investigating georeferenced image organization and summarization [13, 34, 38, 41, 52, 64, 68, 78, 87].

One seminal effort to organize georeferenced photo organization is the World Wide Media eXchange (WWMX) database by Toyama et al. from Microsoft Research [87]. WWMX is a map based prototype system that indexes and browses a large collection of image media with georeference, timestamp, etc. It is the first approach that concretes a number of important issues in building and indexing a geoferenced database, which include: acquisition of georeferences, data structure for georeferences and database optimization for georeferenced media. Thereafter, the map-base photo browsing was adopted by several commercial photo-sharing services on the Internet, such as Flickr and Panaromio. Figure 3 shows the interfaces of WWMX (cited from [87]) and Panaromio. Both WWMX and Panaromio browse photos by super-imposing them on map. Photo overlay on map, however, gives rise to the issue of clutter map, as shown in Fig. 3b.

To alleviate the clutter in the map, several research projects devoted their effort to selecting representative photos [62–64]. PhotoCompas by Naaman et al. [64] attempted to browse personal photo album via a location and event hierarchy, which can facilitate efficient search and browsing for photos of particular events and locations. The location and event hierarchy is constructed via a combination of existing time-based event detection methods [27, 32] and a temporal-geographical clustering algorithm [24]. Specifically, PhotoCompas builds the hierarchical structure in two steps: (1) automatically grouping photos into distinct events and geographical locations, and (2) suggesting intuitive geographical names for the resulting groups

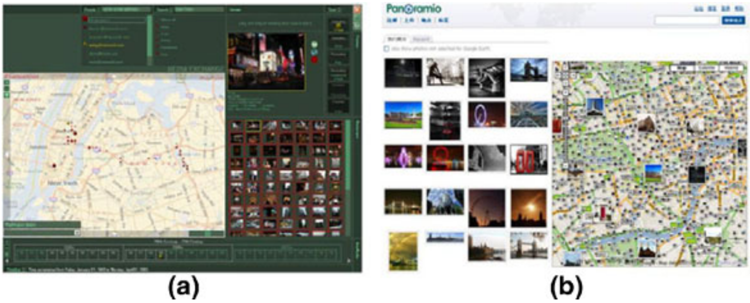

Fig. 3 Interface of WWMX (**a**) (cited from [87]) and Panoramio (**b**). Both WWMX and Panaromio browse photos by super-imposing in a map based system. Photo overlay on map, however, gives rise to the issue of clutter map, as shown in **b**

[64]. The advantage of PhotoCompas is that it allows browsing of photo collection without the use of a map. This is particularly useful in small-screen devices and scenarios where displaying map is not convenient. The user studies in [62, 64] also demonstrated the efficiency and usability of PhotoCompas. In addition to utilizing time and location metadata, an extended version of PhotoCompas was proposed later [63]. The extended PhotoCompas system integrates into the browser's interface a more comprehensive set of location- and time- derived context information, including weather, local time, daylight status, sunset/sunrise time, etc. These context information are extracted from various sources like weather station of the place where the photo is geotagged. A use survey is then conducted to evaluate how effective the context information are in facilitating better search and browsing. Studies show that local time, daylight status, season seem to be stronger cues than weather, temperature, data/time information; and outdoor/indoor cues are not effective or useful in recalling one's memory for effective photo search.

Similar to the approaches in [63, 64], Jaffe et al. [38] developed a system to automatically select representative and relevant photographs from a particular spatial region. The resulting summarization allows users to browse more easily and efficiently through large scale georeferenced photo collections, as shown in Fig. 4. The representative photos of a spatial region are selected by mining photographic behavior patterns from spatial, temporal, and social metadata of photos. Specifically, a modified Hungarian hierarchical clustering [26] is applied to identify groups of spatially adjacent photos. Photos with top ranking scores in the cluster are then

Fig. 4 Representative photos are selected to summarize the photo collection in San Francisco (cited from [38])

selected to be representative ones. Not only summarizing collections with a subset of photos, the system [38] also generates a "Tag Map" to visualize the distribution of textual-topical tags representative to a particular spatial region. In this aspect, the World Explorer system in [4] shares a similar vision. In [4], World Explorer analyzes the tags associated with georeferenced Flickr photos to generate aggregate knowledge in the form of representative tags for arbitrary areas in the world. Metaphorically, it aims to create a "psychological map" of an arbitrary area via location-based information analysis. The analysis is based on multi-level clustering and TF-IDF (term frequenc, inverse document frequency) based scoring of tags. The outcome of World Explorer is to visualize text tags representative to spatial areas on a map, as shown in Fig. 5.

Besides photo metadata, Crandall [18] exploits the local visual features to determine the representative or canonical photos of a specific location. The premise of the approach is based on the collective behavior of photographers. Namely, people take photos because they are visually attracted by the subjects. If more photos are taken on a view, then the view is more attractive and representative. The canonical photo selection is then cast to finding the most salient photo out of a group of visually similar ones. Borrowing the solution of [10, 76], the approach formulates canonical photo selection as a graph problem. In a graph, each node represents a photos and edge indicates the visual similarity of photos. The photos that are most tightly connected to the others are deemed to be the canonical one. In this system, Scale Invariant Feature Transform (SIFT) [54, 55] are used to compute the visual similarity of photos.

3.2 Mining knowledge from georeferenced media

Billions of socially generated media resources on the Internet are a result of experience sharing by web communities. This fast growing media collection records our culture, society and environment, and provides opportunities to mine semantic and social knowledge of this world [35, 38, 44, 70].

3.2.1 Extracting location semantics from geotagged photos

Jaffe et al. [38] and Kennedy et al. [44] from Yahoo! Research first attempted to extract aggregate knowledge on certain location from large scale georeferenced photos at Flickr. The "knowledge" here refers to the word or concept that can best describe and symbolize a geographical region. The challenge is to extract structured

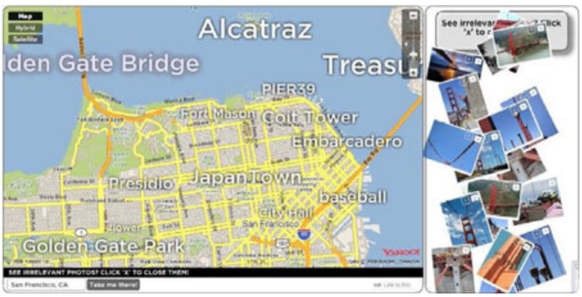

Fig. 5 World Explorer visualizes text tags representative to spatial areas on a map (cited from [4])

knowledge from the unstructure set of tags. The premise of the proposed solution is based on the human attention and behavior embedded in the photos and tags. Namely, if tags concentrate in a geographical area but do not occur often outside that area, then these tags are more representative to the area than those spread over large spatial region. The algorithm is similar to the one in [4], which exploits clustering and TF-IDF to estimate the representativeness of tags.

Rattenbury et al. [71] further investigated the place and event semantics of georeferenced tags, in addition to the representativeness. The proposed approach can automatically determine whether a tag corresponds to a "place" like Bay Bridge or an "event" like F1 car race 2010. A "place" tag is defined as a one that exhibits significant spatial patterns, while an "event" tag refers to a one that exhibits significant temporal patterns. Both definitions are vague and subject to some geographic region. For example, **carnival** may not be able to indicate any event, but will be very specific if only carnivals in New York City are considered. The method, named Scale-structure Identification, is developed to analyze the spatial and temporal distribution of tags and identify the "event" and "place" ones with relative geographic scale. The "event" and "place" semantic identification can be useful to many applications, such as image search, collection browsing and tag visualization [22].

3.2.2 Learning tourism knowledge

In Web 2.0 communities, people share their traveling experience in blogs and forums. As shown in Fig. 6, these articles, named travelogues, contain various tourism related information, including text depiction of landmark, photos of attractions and so on [31, 40, 100]. Travelogue provides abundant data source to extract tourism related knowledge. Hao et al. [31] proposed to exploit travelogues to generate location overviews in the form of both visual and textual descriptions. The approach first mines a set of location-representative keywords from travelogues, and retrieve web images using the learnt keywords. The resulting web images and tags are presented via a user interface to provide an overview for a given location. To model travelogue documents, the approach assumes a document is generated from a mixture of topics. A generative travelogue model is then developed by extending probabilistic latent semantic analysis (pLSA) model [36, 58]. The model learns the word-topic (local and global tourism topic, like an attraction sight) distribution of travelogue documents and identifies representative keywords within a given location.

In later work [30], Hao et al. extended the approach by modeling travelogue documents with a refined model, named Location-Topic Model. Based on travelogue

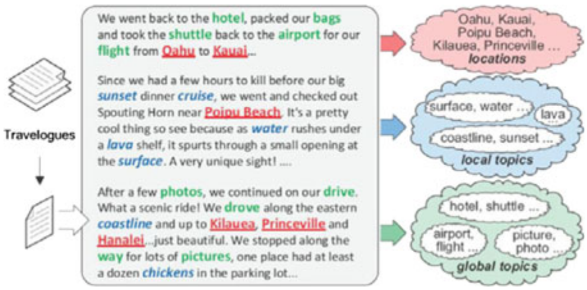

Fig. 6 Example of a travelogue and its topics (cited from [30])

modeling, three applications are further developed, which are: (1) tour destination recommendation, (2) destination summarization, and (3) travelogue enrichment. To recommend a destination, user presents his tour intention by a query like "I plan to go hiking next month. Could you recommend some destinations good for hiking?" [30]. The system then utilizes the topic model to select the destination with highest matching score. The destination summarization visualizes a tour destination with its representative photos and tags, similar to the approach in [31]. To facilitate user to browse other's travelogues, the approach extracts the highlights of a travelogue document and enriches it by providing additional visual descriptions.

Complementing travelogues, georeferenced photos also tell a great deal about tourism knowledge. The photos, together with their time- and geo-references, implicitly document the photographers' spatiotemporal movement paths. Zheng et al. [97] first explored the geotagged photos on Flickr to analyze the people's travel pattern at the local level of a tour destination. First, from a noisy pool of geotagged photos, the approach builds a statistically reliable database of travel paths, and mine a list of regions of attraction (RoA). Then the tourist traffic flow among different RoAs is investigated by exploiting Markov chain model [20]. The Markov chain model is widely used in various disciplines to analyze the trend of spatio-temporal movement and outcomes of sequential events [37, 89]. Based on the first-order dependence in Markov chain, the approach estimates the statistics of visitors traveling from one region to another. Such tourist traffic analysis helps to indicate centric regions of attractions (RoA), which have influx of tourists from many other RoAs. Testing is conducted on four major cities, including San Francisco, New York City, Paris and London, and demonstrates encouraging and interesting results.

Before the emergence of geotagged photos on the Internet, people mobility and travel behavior within a local tour destination have been actively researched, as they are important topics to mobile applications and location based services [6, 49, 56, 57, 97]. In general, there exist two types of methods to acquire detailed travel data: (1) a survey with questionnaire on people's location histories [57]; and (2) location-acquistion devices for people to wear, such as GPS, cellular phone, etc. [57, 101]. The issue with the first method is its expensive and time-consuming manual process, while the second method gives rise to unavoidable privacy issue that makes most people reluctant to participate in the study. The approach in [97] circumvents these two issues by acquiring people's travel information from GPS-tagged photos on the Internet. The advantage of this approach is that tourist mobility analysis can readily scale up onto a multitude of tour destinations. Such an automated travel pattern analytic approach can be tremendously useful to many geo-spatial applications. For example, the travel sequence analysis can reveal the crowd's choice of popular tour routes and help to monitor the traffic patterns of tourists [97].

3.2.3 Culture discovery from geotagged photos

The georeferenced photos and tags contain rich information about the culture of their geotagged region. Yanai et al. [93–95] attempted to detect the cultural differences of certain local regions by mining the representative photos of selected culture concepts. Given a concept, such as "noodle", the approach first locates a set of relevant photos across different geographic areas. Then the geographic regions representative to the concept are identified via a clustering approach. The rational is that if a geographic region contains a multitude of concept relevant photos, then the region is representative to the concept. Finally, for each region, a set of representative photos are selected to visualize the culture concepts in the region. Figure 7 illustrates the representative photos and regions for concept "noodle" in Japan and Europe. In essence, the approach [93–95] is to identify representative photos and regions in a image collection, similar to previous work [4, 38, 64]. The difference is that the focus here is on detecting cultural differences of particular regions.

3.3 Learning landmarks in the world

Landmark is a prominent geographic feature that exhibits salient visual, cultural, natural, functional or historical significance. The attractiveness and popular appeal of landmarks result in a vast amount of landmark related media resources on the Internet, which has spurred much research attention from SIGGRAPH, Computer Vision and Multimedia communities. This section reviews the literature work on landmark from three aspects: (1) building world-wide landmark database, (2) landmark visual summarization and 3D modeling, and (3) landmark recognition.

3.3.1 Building world-wide landmark database

Several studies have investigated on building landmark database [1, 43, 50]. The scale of constructed databases is usually at city level. Zheng et al. [98, 99] first attempted to construct a world-scale landmark database, including landmarks' photos, country, city, GPS latitude and longitude, and so on. The approach first mines a comprehensive list of landmarks from two sources: (1) 20 million geotagged photos and (2) online tour guide web pages. Candidate images for each landmark are then obtained from photo sharing websites or by querying an image search engine. Second, landmark visual models are built by pruning candidate images using efficient image matching and unsupervised clustering techniques. Finally, the landmarks and their visual models are validated by checking authorship of their member images. The resulting landmark database incorporates 5,312 landmarks from 1,259 cities in 144 countries. Figure 8 shows the distribution of the landmarks in the database.

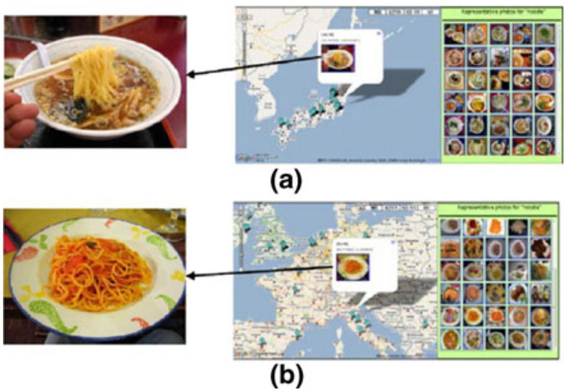

Fig. 7 Representative photos and regions for concept "noodle" in Japan (**a**) and Europe (**b**) (cited from [93])

Fig. 8 The distribution of landmarks mined from geotagged photos and tour guide web articles (cited from [98])

3.3.2 Landmark visual summarization and 3D modeling

As landmark photos on the Internet grow rapidly, browsing and summarizing landmark photo collections become important. Kennedy et al. [43] proposed to extract representative views for a landmark by unsupervised learning on Flickr photos and metadata. Geographic tags corresponding to landmark names are first identified via a clustering approach similar to [4]. The visual contents of photos associated with landmark tags are then summarized to generate representative views for the corresponding landmark. The basis is that the landmark view is deemed to representative, if many people photograph it. Similarly, Chen et al. [15] summarized a landmark with a set of canonical and consensus images via image matching and clustering. The resulting canonical image is then segmented into landmark icon, which is further incorporated into a tour map.

Parallelled by summarizing landmarks with representative photos [15, 43, 72], researchers from SIGGRAPH and Computer Vision communities also explore how to summarize and browse landmark photos in a 3D environment. *Photo Tourism* by Microsoft Research [79, 80] first explored the sparse 3D reconstruction of landmarks from Internet photos. The system achieves high quality reconstructions, by exhaustive pairwise image matching and global bundle adjustments of the model after inserting each new view into the 3D model [50]. The major limitation is, however, the high computational complexity and sensitivity to outlier images. The process becomes particularly inefficient, when the photo collection is large and heavily contaminated with noisy non-landmark images. Based on the Photo Tourism system, Microsoft developed a commercial software, i.e. Photosynth. To circumvent the issues of efficiency and outlier images, Photosynth accepts a set of predominantly "clean" landmark photos from user as input. Inspired by Photo Tourism, more efficient structure from motion (SfM) methods are developed to perform landmark 3D construction [2, 3, 25, 50, 65, 81]. The basis is similar, which is utilizing the visual redundancy of comuunity-contributed photos to learn the appearance and 3D geometric structures of landmarks.

3.3.3 Landmark recognition

Along the development of landmark databases, many landmark recognition systems have been proposed [50, 51, 98]. Despite the differences in scale and methodologies, most approaches formulate landmark recognition as a classification task. Different from traditional classification tasks, like object categorization, landmark recognition encompasses challenges of a much larger number of landmark categories and a vast amount of landmark images. For example, the recognition engine in [98] incorporates over 5,000 landmark categories and about 1 million images in the model. Thus, efficiency becomes a non-trivial issue. Moreover, landmark photos may present huge photometric and geometric variations, due to changes in scale, pose, translation, image capturing conditions, viewpoint, occlusion and clutter [48].

A reliable image representation is crucial to build effective visual models of landmarks. Global features, such as color moments, color correlogram, are sensitive to changes in scale, pose, illumination and image capturing condition. On the other hand, part-based local image representation, like bag of local features, have shown robustness and resilience in photometric and geometric image variations, such as changes in scale, translation, light condition, viewpoint, occlusion and clutter, in part [48, 53]. Thus, most landmark recognition systems [50, 51, 98] adopt local feature representation by extracting a set of informative and highly repeatable interest points (regions), based on color or geometric saliency. Examples of local region detectors are Difference of Gaussian [55], Harris–Laplace [59], Maximally Stable Extremal Regions (MSRE) [21], to name a few. For each detected keypoint, a feature descriptor (vector) is computed over its local neighborhood. There exist several local descriptors, such as Gradient Location and Orientation Histogram (GLOH) [60], Scale Invariant Feature Transform (SIFT) [54, 55], Speeded Up Robust Features (SURF) [7], Shape Context [8], Spin Image [47], and so on. By representing images as a bag of unordered local features, the pairwise image similarity can then be estimated via a image matching model [55]. To quantitatively determine the match score between two images, the recognition system in [98] models the local feature matching as a stochastic Bernoulli process [73]. In the process, the outcome of each feature matching is regarded as a Bernoulli random variable that are identically and independently distributed (i.i.d). The match score is estimated by using a cumulative binomial distribution, in the spirits of [69]. To achieve high efficiency, a indexing mechanism, like k-d tree [9], is usually used to retrieval nearest neighbors of local features, in the phase of image matching.

In addition to bag of features representations, the recognition systems in [50, 51] adopt bag-of-words and its variations to represent images. The advantage is obvious. The vector representation allows most distance metric and discriminative classifiers, such as the nearest neighbor schemes and the kernel based classifiers, to be readily applicable to landmark recognition. To date, landmark recognition has achieved substantial progresses. According to [98], the recognition performance on over 5,000 landmarks reaches an accuracy of 80.8%; and the time it takes to recognize landmark in a query images is only 0.2 s in a P4 computer. The landmark recognition system in [98] has been incorporated into the Google's newly released mobile product, Goggles.[8]

3.4 Estimating geographic location of a photo

While geographic metadata of photos have been actively studied, research attention is also drawn to the other end of the spectrum: recognizing the geographic location

[8] http://www.google.com/mobile/goggles/

of a photo. The "Where am I" Contest held at ICCV'05 [84] has provided a platform to showcase the state-of-the-arts visual localization methods [28, 96]. In the contest, participants are given a collection of photographs in a city along with the GPS location; and the task is to estimate the geographic location of an unseen image. Multi-view geometric feature based image matching is the basis for most participating approaches [28, 96].

Soon after the contest, the gigantic geotagged photo collections available on Flickr fueled the visual localization towards world scale. The IM2GPS system by Hays and Efros [33] estimates the geographic location of a photo in a purely data-driven scene matching approach. Given an unseen photo, the approach retrieves top 120 most visually similar photos out of a pool of 6 million geotagged photos. The probability distribution of geographic location is estimated from the weighted GPS coordinates of the 120 photos. The mode of the distribution is determined using mean-shift clustering and used as the prediction of photo location. Figure 9 shows examples of query images and their estimated geographic location probability on the map. Kalogerakis et al. [42] extended IM2GPS system to identify geographic location for sequences of timestamped photos. By utilizing 6 million geotagged photos from Flickr as training database, a travel prior distribution is estimated to describe the likelihood of traveling from one location to another during a given time interval. Similar to [33], the geographic location distribution of an input image is determined by matching it against the photos in the training database. The locations for images in a sequence is then inferred by using the Forward-Backward algorithm [12].

The methodologies of [33, 42] enable them to provide generic geographic location estimation on photos taken anywhere in the world in theory. The price is that the location prediction accuracy may not reach the level of satisfaction, as existing geotagged photos are not sufficient to provide extensive visual sampling of the planet Earth. On the other hand, some researchers choose to perform relatively accurate visual localization on a limited set of geographic locations, namely tourist landmarks [15, 18, 50, 51, 98]. The reason is obvious. The popular appeal of landmarks always attracts people's attention, and consequently, the sheer volume of photos provide extensive visual samples of landmark appearances and geometric structures. Refer to

Fig. 9 Given query images (in the *left*), their geographic location probability on the map (in the *right*) are estimated from the top *k* nearest neighbor geotagged photos (in the *middle*) in the training database (cited from [33])

Section 3.3.3 for the survey of literature work on landmark recognition. The reported accuracy of landmark recognition is much higher than the general photo location estimation.

Besides visual features, researchers also exploited text metadata to estimate the geographic location of photos. Serdyukov et al. [75] explored a language model on Flickr photo tags to predict the geographic location of photos. The approach discretizes the earth's surface into $m \times n$ grids, each of which defines a location. A multinomial language model is then estimated from the tags associated with images in the grid for location prediction. Similarly, Laere et al. [91] proposed to geotag Flickr photos by analyzing the distributions of photo tags. The approach first clusters regions of interest into disjoint areas and compiles a vocabulary of relevant text tags using χ^2 statistic. A Naive Byes classifier is then trained on the tag vocabulary to predict geographic area of unseen photos.

It is worth mentioning that place recognition has been a well studied research topic in robotics, even before the emergence of geotagged photos on the Internet. In the context of mobile robot system, place recognition (or robot localization/mapping) is crucial to robot navigation, as it determines and tracks the position of a mobile robot relative to its environment [46, 74, 86, 88]. The different context and target make researchers approach robot localization and location recognition in photos in disparate methodologies. The place recognition in robotics is usually to identify the robot location within a constrained environment like an office building, while location recognition in photos in recent research targets to estimate the general geographic location where the photo was shot. Nevertheless, many techniques in visual robot localization, such as visual features [96], have been borrowed and extended to location recognition in photos [28, 96]. Refer to [19] for survey on robot localization and navigation.

4 Research challenges and future directions

Research on georeferenced media has achieved many advances in various aspects. However, there are still many open research issues that need to be solved to build compelling geographic and location based applications.

4.1 Multilingual mining in georeferenced media

As introduced earlier, georeferenced media on the Internet is collaboratively authored and shared by web community across the world. Georeferenced media is, therefore, multilingual in nature. However, most systems take English as the processing language only. This effectively excludes the media resources in other languages. The consequence is that the knowledge and patterns mined from georeferenced media are biased towards English speaking countries and regions. The study in [98] confirms this conjecture. In [98], a world-scale landmark database is built from geotagged photos and tour guide articles. Observation shows that among top 20 countries with largest number of landmarks, 17 of them are in North America or Europe where English is commonly spoken. In particular, the number of landmarks in China amounts to 101 only, which is clearly under-counted. This manifests the need to incorporate multilingual language processing into data mining on georeferenced media.

Community-contributed media resources on the Internet are the result of experience sharing by Web 2.0 users. Media resources in different languages may reflect the knowledge, vision and perception of different cultures and community. Leveraging geographic location and language in an inter-connected fashion opens up possibilities to learn different behavioral and social patterns in different cultures.

4.2 Geographic orientation of photos

The geographic locations of photos on the Internet have opened up a new host of research and application possibilities. Knowing the geograhic orientation of photos, i.e., in which direction the cameras are pointing, will broaden the opportunities even further. Though most cameras are not equipped with sensors to measure the orientation and inclination of the device, smartphotos, with the iPhone and HTC Magic as prime examples, have started to embrace digital compass technologies [17]. In addition to hardware sensors, software solutions to estimate photo orientation also exist [16, 79]. For example, the Photo Tourism system [79] estimates the relative translation and orientation between photos, by leveraging the visual redundancy among photos. Till now, geographic orientation of photos are rarely available. Nevertheless, with the development of compass-equipped cameras and smartphones, such kind of metadata is expected to emerge in the near future. With the availability of photo orientation metadata, many compelling applications can be accomplished. For example, with the photo alignment information, visual summarization and browsing of photo collections can be adaptive to the user direction and perspective on the map. Moreover, 3D reconstruction of geo-location can be much more efficient.

4.3 Travel guide from geotagged photos

Georeferenced media resources, like tour guide articles and travelogues, contain rich information about tourism. While much research effort has been invested in digging knowledge and patterns of touristic attractions [31, 40, 100], little has been done on how to travel among these attractions. Designing a travel guide is a much more complicated task that need to takes into account not only itinerary, traveling sequence and staying time of different sights, but also personal preferences. Recently, researchers from Yahoo [90] developed a travel guide system on popular locations and itineraries from geotagged Flickr photos. The premise of the system is to utilize the "wisdom of the crowd" via a data-driven approach. Nevertheless, many issues remain open and need to be further explored.

4.4 Social tagging of locations and events

Recent popularity of location-based social services, such as the Foursquare,[9] Gowalla,[10] and Hot-Potato,[11] etc., have generated huge amount of detailed location and event tags. It covers not only popular landmarks, but also obscure places, thus

[9] http://foursqure.com

[10] http://gowalla.com/

[11] http://hotpot.uvic.ca/

providing broad and wide coverage of locations in unprecedented scales. Integration of these social location tags and geotagged photos permits not just popular locations to be recognized, but also little known places like student hostels in universities, grocery stores in neighborhood and so on. It opens up new opportunity to expand research and applications from popular (or mostly touristic) places to obscure locations.

5 Concluding remarks

Research on online georeferenced multimedia is a young field dating back to the early 2000's, yet it has been actively studied in several major research communities of Computer Vision, Multimedia, KDD and Digital Libraries. This paper surveyed the recent research and applications on georeferenced media to highlight what has been achieved and what are the challenges and possible research directions. Specifically, the survey focuses on the following four aspects: (1) organizing and browsing georeferenced media resources, (2) mining semantic/social knowledge from georeferenced media, (3) learning landmarks in the world, and (4) estimating geographic location of a photo. Based on the research achievements now, open issues and challenges are identified, and directions that can lead to compelling applications are suggested.

References

1. Abbasi R, Chernov S, Nejdl W, Paiu R, Staab S (2009) Exploiting flickr tags and groups for finding landmark photos. In: Advances in information retrieval. Lecture notes in computer science, vol 5478. Springer, Berlin, pp 654–661
2. Agarwal A, Furukawa Y, Snavely N, Curless B, Seitz SM, Szeliski R (2010) Reconstructing Rome. Computer 43:40–47
3. Agarwal S, Snavely N, Simon I, Seitz SM, Szeliski R (2009) Building Rome in a day. In: Proceedings of international conference on computer vision. Kyoto, Japan
4. Ahern S, Naaman M, Nair R, Yang JH-I (2007) World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries. ACM, New York, pp 1–10
5. Anselin L (1992) Spatial data analysis with GIS: an introduction to application in the social sciences. Technical report, University of California, Santa Barbara
6. Asakura Y, Iryo T (2007) Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transp Res, Part A Policy Pract 41(7):684–690
7. Bay H, Tuytelaars T, Gool V, Surf L (2006) Speeded up robust features. In: 9th European conference on computer vision. Graz Austria
8. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
9. Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18(9):509–517
10. Berg TL, Forsyth D (2007) Automatic ranking of iconic images. Technical report UCB/EECS-2007-13, EECS Department, University of California, Berkeley
11. Berry JK (1993) Beyond mapping: concepts, algorithms, and issues in GIS/Joseph K. Berry. GIS World, Inc., Ft. Collins, Colo., USA
12. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

13. Camara AS, Raper J (eds) (1999) Spatial multimedia and virtual reality. Taylor & Francis, Bristol
14. Cayzer S, Butler MH (2004) Semantic photos. Technical report, HP Laboratories Bristol
15. Chen W-C, Battestini A, Gelfand N, Setlur V (2009) Visual summaries of popular landmarks from community photo collections. In: MM '09: Proceedings of the seventeen ACM international conference on multimedia. ACM, New York, pp 789–792
16. Chippendale P, Zanin M, Andreatta C. Visual environment monitoring: the marmota project. http://tev.fbk.eu/marmota/
17. Chippendale P, Zanin M, Andreatta C (2009) Collective photography. In: Conference for visual media production, pp 188–194
18. Crandall DJ, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 761–770
19. DeSouza GN, Kak AC (2002) Vision for mobile robot navigation: a survey. IEEE Trans Pattern Anal Mach Intell 24:237–267
20. Diaconis P (2009) The Markov chain Monte Carlo revolution. Bull Am Math Soc New Ser 46(2):179–205
21. Donoser M, Bischof H (2006) Efficient maximally stable extremal region (MSER) tracking. In: Proceedings of conference on computer vision and pattern recognition, pp 553–560
22. Dubinko M, Kumar R, Magnani J, Novak J, Raghavan P, Tomkins A (2006) Visualizing tags over time. In: Proceedings of the 15th international conference on World Wide Web. ACM, New York, pp 193–202
23. Ester M, Kriegel HP (1997) Spatial data mining: a database approach. In: Advances in spatial databases. Springer, Berlin, pp 47–66
24. Gionis A, Mannila H (2003) Finding recurrent sources in sequences. In: Proceedings of the annual international conference on research in computational molecular biology. ACM, New York, pp 123–130
25. Goesele M, Snavely N, Curless B, Hoppe H, Seitz SM (2007) Multi-view stereo for community photo collections. In: Proceedings of IEEE conference on computer vision, Rio de Janeiro, Brazil, 14–20 October 2007
26. Goldberger J, Tassa T (2008) A hierarchical clustering algorithm based on the Hungarian method. Pattern Recogn Lett 29(11):1632–1638
27. Graham A, Garcia-Molina H, Paepcke A, Winograd T (2002) Time as essence for photo browsing through personal digital libraries. In: Proceedings of the ACM/IEEE-CS joint conference on digital libraries. ACM, New York, pp 326–335
28. Hakeem A, Vezzani R, Shah M, Cucchiara R (2006) Estimating geospatial trajectory of a moving camera. In: ICPR '06: Proceedings of the 18th international conference on pattern recognition. IEEE Computer Society, Washington, DC, pp 82–87
29. Han J, Kamber M, Tung AKH (2001) Geographic data mining and knowledge discovery, chapter. Spatial clustering methods in data mining: a survey. Taylor and Francis
30. Hao Q, Cai R, Wang C, Xiao R, Yang J-M, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: WWW '10: Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 401–410
31. Hao Q, Cai R, Wang X-J, Yang J-M, Pang Y, Zhang L (2009) Generating location overviews with images and tags by mining user-generated travelogues. In: Proceedings of the seventeen ACM international conference on multimedia. ACM, New York, pp 801–804
32. Harada S, Naaman M, Song YJ, Wang QY, Paepcke A (2004) Lost in memories: interacting with photo collections on PDAS. In: JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries. ACM, New York, pp 325–333
33. Hays J, Efros A (2008) IM2GPS: estimating geographic information from a single image. In: Proceedngs of conference on computer vision and pattern recognition
34. Heuer JT, Dupke S (2007) Towards a spatial search engine using geotags. In: Kessler C, Probst F (eds) GI-Days 2007—young researchers conference. Institute for Geoinformatics, pp 199–204. citeulike-article-id=1668715
35. Hile H, Vedantham R, Cuellar G, Liu A, Gelfand N, Grzeszczuk R, Borriello G (2008) Landmark-based pedestrian navigation from collections of geotagged photos. In: Proceedings of the 7th international conference on mobile and ubiquitous multimedia. ACM, New York, pp 145–152

36. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of uncertainty in artificial intelligence. UAI, Stockholm
37. Ishikawa Y, Tsukamoto Y, Kitagawa H (2004) Extracting mobility statistics from indexed spatio-temporal datasets. In: Spatio-temporal database management, 2nd international workshop STDBM'04, Toronto, Canada, 30 August 2004, pp 9–16
38. Jaffe A, Naaman M, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval. ACM, New York, pp 89–98
39. Jesdanun A (2008) GPS adds dimension to online photos citation. In: ABC news, technology & science. http://www.physorg.com/news119889687.html. Accessed 18 Jan 2008
40. Jing F, Zhang L, Ma W-Y (2006) Virtualtour: an online travel assistant based on high quality images. In: Proceedings of the 14th annual ACM international conference on multimedia. ACM, New York, pp 599–602
41. Jung V (1999) Metaviz: visual interaction with geospatial digital libraries. Technical report, 4] INVISIP—Information Visualisation for Site Planning
42. Kalogerakis E, Vesselova O, Hays J, Efros AA, Hertzmann A Image sequence geolocation with human travel priors. In: Proceedings of international conference on computer vision. Kyoto, Japan
43. Kennedy L, Naaman M (2008) Generating diverse and representative image search results for landmarks. In: Proceeding of the 17th international conference on World Wide Web. ACM, New York, pp 297–306
44. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T (2007) How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of conference on multimedia. ACM, New York, pp 631–640
45. Koperski K, Adhikary J, Han J (1996) Spatial data mining: progress and challenges survey paper. In: SIGMOD workshop on research issues on data mining and knowledge discovery, pp 1–10
46. Kuipers B, Beeson P (2002) Bootstrap learning for place recognition. In: Eighteenth national conference on artificial intelligence. American Association for Artificial Intelligence, Menlo Park, pp 174–180
47. Lazebnik S, Schmid C, Ponce J (2005) A sparse texture representation using local affine regions. IEEE Trans Pattern Anal Mach Intell 27(8):1265–1278
48. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of conference on computer vision and pattern recognition, pp 2169–2178. Washington, DC, USA
49. Lewa A, McKerchera B (2006) Modeling tourist movements: a local destination analysis. Ann Tour Res 33(2):403–423
50. Li X, Wu C, Zach C, Lazebnik S, Frahm J-M (2008) Modeling and recognition of landmark image collections using iconic scene graphs. In: Proceedings of European conference on computer vision, pp 427–440
51. Li Y, Crandall DJ, Huttenlocher DP (2009) Landmark classification in large-scale image collections. In: Proceedings of international conference on computer vision, pp 1957–1964. Kyoto, Japan
52. Lim E-P, Goh DH-L, Ng ZLW-K, Liu Z, Ng WK, Khoo CSG, Higgins SE (2002) G-portal: a map-based digital library for distributed geospatial and georeferenced resources. In: Proceedings of the second ACM+IEEE joint conference on digital libraries, pp 351–358
53. Lisin DA, Mattar MA, Blaschko MB, Learned-Miller EG, Benfield MC (2005) Combining local and global image features for object class recognition. In: CVPR '05: Proceedings of the 2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05)—workshops. IEEE Computer Society, Washington, DC, p 47
54. Lowe DG (1999) Object recognition from local scale-invariant features. In: IEEE international conference on computer vision, vol 2, pp 1150–1157
55. Lowe DG (2003) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 20:91–110
56. McKercher B, Lau G (2007) Understanding tourist movement patterns in a destination: a GIS approach. Tour Hosp Res 7(1):39–49
57. Mckercher B, Lau G (2008) Movement patterns of tourists within a destination. Tour Geogr 10(3):355–374

58. Mei Q, Liu C, Su H, Zhai CX (2006) A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: WWW '06: Proceedings of the 15th international conference on World Wide Web. ACM, New York, pp 533–542
59. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
60. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
61. Miller HJ, Han J (2001) Geographic data mining and knowledge discovery. Taylor & Francis, Bristol
62. Naaman M, Harada S, Wang Q, Paepcke A (2004) Adventures in space and time browsing personal collection of geo-referenced digital library. Technical report, Stanford University
63. Naaman M, Harada S, Wang QY, Garcia-Molina H, Paepcke A (2004) Context data in geo-referenced digital photo collections. In: Proceedings of the ACM international conference on multimedia. ACM, New York, pp 196–203
64. Naaman M, Song YJ, Paepcke A, Garcia-Molina H (2004) Automatic organization for digital photographs with geographic coordinates. In: JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries. ACM, New York, pp 53–62
65. Ni K, Steedlyy D, Dellaert F (2007) Out-of-core bundle adjustment for large-scale 3d reconstruction. In: Proceeding of international conference on computer vision, Rio de Janeiro, Brazil, 14–20 October 2007
66. Niculescu D, Nath B (2001) Ad hoc positioning system (APS). In: Globecom, pp 2926–2931
67. Panoramio blog. http://www.panoramio.com/blog/1-million-geolocated-photos-at-panoramio/
68. Pigeau A, Gelgon M (2004) Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. J Vis Commun Image Represent 15(3):425–445
69. Pope AR, Lowe DG (2000) Probabilistic models of appearance for 3-d object recognition. Int J Comput Vis 40(2):149–167
70. Quack T, Leibe B, Gool LV (2008) World-scale mining of objects and events from community photo collections. In: CIVR '08: Proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, pp 47–56
71. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of ACM SIGIR. ACM, New York, pp 103–110
72. Ren Y, Yu M, Wang X-J, Zhang L, Ma W-Y (2010) Diversifying landmark image search results by learning interested views from community photos. In: Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 1289–1292
73. Rota G-C, Baclawski K (1979) Introduction to probability and random processes
74. Se S, Lowe D, Little J (2001) Vision-based mobile robot localization and mapping using scale-invariant features. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 2051–2058
75. Serdyukov P, Murdock V, van Zwol R (2009) Placing flickr photos on a map. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 484–491
76. Simon I, Snavely N, Seitz SM (2007) Scene summarization for online image collections. In: Proceedings of international conference on computer vision, Kyoto, Japan. IEEE, pp 1–8
77. Skyhook. http://www.skyhookwireless.com/
78. Smith TR (1996) A digital library for geographically referenced materials. Computer 29(5): 54–60
79. Snavely N, Seitz SM, Szeliski R (2006) Photo tourism: exploring photo collections in 3d. In: ACM transactions on graphics. ACM, New York, pp 835–846
80. Snavely N, Seitz SM, Szeliski R (2008) Modeling the world from Internet photo collections. Int J Comput Vis 80(2):189–210
81. Snavely N, Seitz SM, Szeliski R (2008) Skeletal sets for efficient structure from motion. In: Proceeding of conference on computer vision and pattern recognition. Anchorage, Alaska, USA
82. Spinellis DD (2003) Position-annotated photographs: a geotemporal web. IEEE Pervasive Computing 2(2):72–79
83. Srinivasan A, Richards JA (1993) Analysis of GIS spatial data using knowledge-based methods. Int J Geogr Inf Syst 7(6):479–500

84. Szeliski R (2009) “Where am i?”: ICCV 2005 computer vision contest. In: Proceedings of the seventeen ACM international conference on multimedia. ACM, New York, pp 961–962
85. Torniai C, Battle S, Cayzer S (2007) Sharing, discovering and browsing geotagged pictures on the web. Technical report, HP Laboratories Bristol, 15 May 2007
86. Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: Proceedings of the ninth IEEE international conference on computer vision. IEEE Computer Society, Washington, DC, p 273
87. Toyama K, Logan R, Roseway A (2003) Geographic location tags on digital images. In: Proceedings of the ACM international conference on multimedia. ACM, New York, pp 156–166
88. Ulrich W, Nourbakhsh I (2000) Appearance-based place recognition and I. In: Proceedings of IEEE international conference on robotics and automation, San Francisco, CA, pp 1023–1029
89. Upton GJG, Fingleton B (1989) Spatial data analysis by example. Vol. 2: Categorical and directional data. Wiley, New York
90. Valentino-DeVries J (2010) Using flickr photos as a travel guide. Wall Street J July 23. http://blogs.wsj.com/digits/2010/07/23/using-flickr-photos-as-a-travel-guide/
91. Van Laere O, Schockaert S, Dhoedt B (2010) Towards automated georeferencing of flickr photos. In: Proceedings of the 6th workshop on geographic information retrieval. ACM, New York, pp 1–7
92. Wagenaar WA (1986) My memory: a study of autobiographical memory over six years. Cogn Psychol 18:225–252
93. Yanai K, Kawakubo H, Qiu B (2009) A visual analysis of the relationship between word concepts and geographical locations. In: Proceeding of the ACM international conference on image and video retrieval. ACM, New York, pp 1–8
94. Yanai K, Qiu B (2009) Mining cultural differences from a large number of geotagged photos. In: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 1173–1174
95. Yanai K, Yaegashi K, Qiu B (2009) Detecting cultural differences using consumer-generated geotagged photos. In: Proceedings of the 2nd international workshop on location and the web. ACM, New York, pp 1–4
96. Zhang W, Kosecka J (2006) Image based localization in urban environments. In: Proceedings of the third international symposium on 3D data processing, visualization, and transmission (3DPVT’06). IEEE Computer Society, Washington, DC, pp 33–40
97. Zheng Y-T, Li Y, Zha Z-J, Chua T-S (2011) Mining travel patterns from GPS-tagged photos. In: Proceedings of ACM conference on multimedia modeling, Taipei, Taiwan, 5–7 Jan 2011. ACM, New York
98. Zheng Y-T, Zhao M, Song Y, Adam H, Buddemeier U, Bissacco A, Brucher F, Chua T-S, Neven H (2009) Tour the world: building a web-scale landmark recognition engine. In: Proceedings of international conference on computer vision and pattern recognition, Miami, FL, USA
99. Zheng Y-T, Zhao M, Song Y, Adam H, Buddemeier U, Bissacco A, Brucher F, Chua T-S, Neven H, Yagnik J (2009) Tour the world: a technical demonstration of a web-scale landmark recognition engine. In: Proceedings of the seventeen ACM international conference on multimedia. ACM, New York, pp 961–962
100. Zheng Y, Zhang L, Xie X, Ma W-Y (2009) Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 791–800
101. Zheng Y, Zhang L, Xie X, Ma W-Y (2009) Mining interesting locations and travel sequences from GPS trajectories. In: WWW ’09: Proceedings of the 18th international conference on World Wide Web. ACM, New York, pp 791–800

Yan-Tao Zheng is a research engineer at Institute for Infocomm Research (I2R), Singapore. He received his Ph.D. from National University of Singapore and B.Eng. (with 1st class honors) from Nanyang Technological University, Singapore. His research interests include geo-mining in multimedia, image annotation and video search. He is the recipient of a number of international awards, including Champion of Star Challenge, Microsoft Research Fellowship, IBM Waston Emerging Multimedia Leaders, Best Ph.D. Thesis Award and so on. During his attachment at Google Inc. in 2008, he developed a world-scale landmark recognition engine together with Google engineers, which has been highly praised and well publicized. He has served as program committee member and reviewer of a number of prestigious international conferences and journals.

Zheng-Jun Zha is a Postdoctoral Research Fellow in School of Computing, National University of Singapore (NUS). He received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. His current research interests include large-scale media search, social media sharing and management, computer vision and machine learning. He received Microsoft Fellowship in 2007 and President Scholarship of Chinese Academy of Science in 2009.

Tat-Seng Chua is the KITHC Chair Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School during 1998–2000. Dr. Chua's main research interest is in multimedia information retrieval, in particular, on the extraction, retrieval and question-answering (QA) of text, video and live media. He is currently working on several multi-million-dollar projects: interactive media search, local contextual search, and live media search. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations. Dr. Chua is active in the international research community. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the conference co-chair of ACM Multimedia 2005, ACM CIVR 2005, and ACM SIGIR 2008. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is the member of steering committee of ICMR (International Conference on Multimedia Retrieval) and Multimedia Modeling conference series; and as member of International Review Panel of two large-scale research projects in Europe. Dr Chua is leading a joint Center between NUS and Tsinghua University to develop technologies for live media search.