Geotagging in Multimedia and Computer Vision – A Survey

Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher
Kodak Research Laboratories, Eastman Kodak Company

Abstract

Geo-tagging is a fast-emerging trend in digital photography and community photo sharing. The presence of geographically relevant metadata with images and videos has opened up interesting research avenues within the multimedia and computer vision domains. In this paper, we survey geo-tagging related research within the context of multimedia and along three dimensions: 1) Modalities in which geographical information can be extracted, 2) Applications that can benefit from the use of geographical information, and 3) The interplay between modalities and applications. Our survey will introduce research problems and discuss significant approaches. We will discuss the nature of different modalities and lay out factors that are expected to govern the choices with respect to multimedia and vision applications. Finally, we discuss future research directions in this field.

1. Introduction

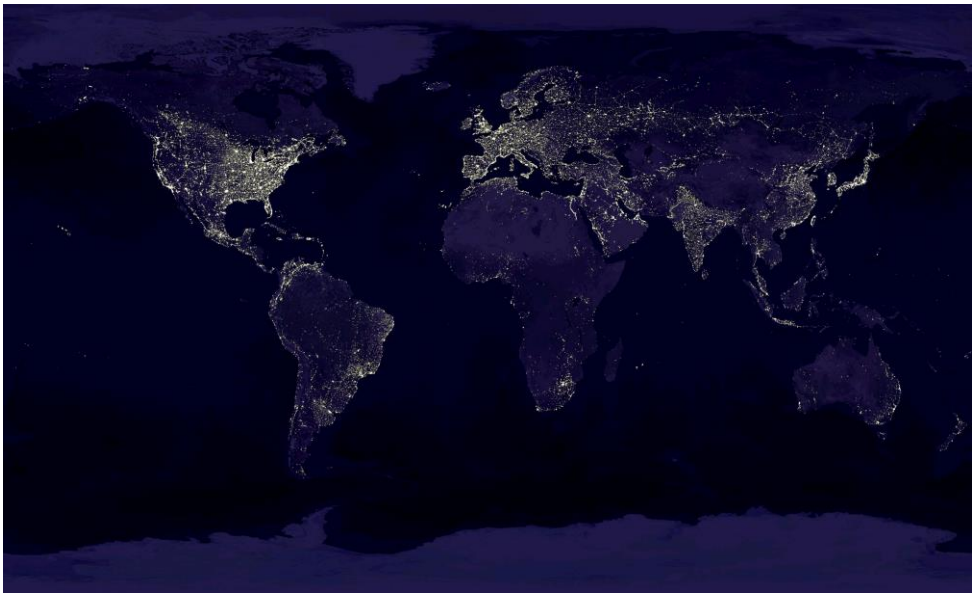
We live in an information age touched by technology in all aspects of our existence, be it work, entertainment, travel, or communication. The extent to which information pervades our lives today is evident in the growing size of personal and community footprints on the web, ever improving modes of communication, and fast evolving internet communities (such as Flickr, Twitter, and Facebook) promoting virtual interactions. In some aspects, man has transformed from a social being into an e-social being. Images and video constitute a huge proportion of the Web information that is being added or exchanged every second. The popularity of digital cameras and camera phones has contributed to this explosion of personal and Web multimedia data. To motivate our readers, in Fig. 1, we visualize two different profiles of human presence or activity in the world; (a) *The World at Night*, captured aerially at night, where the lit areas represent human habitation and activities, and (b) *The Photographed World*, constructed using 1.2 million Flickr pictures, where the bright areas represent density of photographic activity.

Readers should note that we do not superimpose a world map in Fig. 1 and it is the human eye that traces out the continents and coast lines. These boundaries are real in Fig. 1(a) where the Earth is viewed aerially but purely virtual in Fig. 1(b). Notice the similarity between the two maps making one jump to the conclusion that the more populated regions usually invite greater levels of photographic activity. While this is true in general, several pockets of high intensity in the map at the bottom suggest differences between inhabitation and picture-taking patterns perhaps in deserts, reserves, and national parks that have far more visitors than inhabitants. The former is just one example of a useful inference that can be drawn from a large amount of data. It is important to point out that mapping of the world pictures (Fig. 1(b)) would not have been possible if it were not for the people who geotagged their pictures. Therefore this paper is a tribute to geotagging and is devoted to surveying how this phenomenon is changing the face of multimedia research.

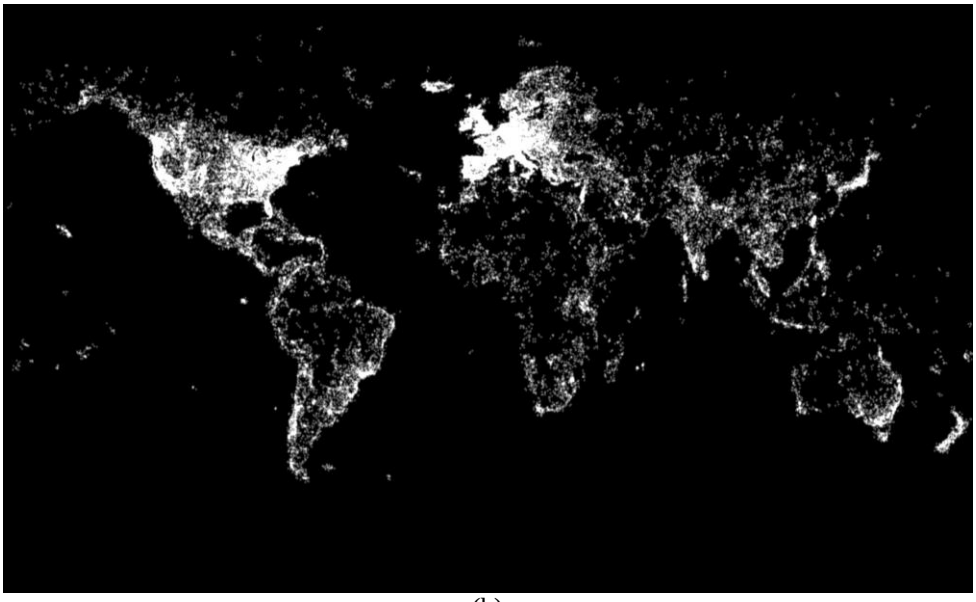
Geotagging or geo-referencing is the process of adding geographical identification metadata to various media such as images and videos in websites, blogs, or photo-sharing web-services [2,4]. It can help users find a wide variety of location-specific information. For example, one can find images taken near a given location by entering latitude and longitude coordinates into a geotagging-enabled image search engine or just by clicking on a region in Google Map. Geotagging-enabled information services [81] can also potentially be used to find location-based news, websites, or other resources. Associating time and place with pictures has always been natural for people. In the past, this association was manifested in more tangible forms such as writing the date and place where the picture was taken on the back of the print. Geotagging has generated a wave of geo-awareness in multimedia repositories and research communities alike[1][2]. Yahoo Flickr has amassed about 4.7 million images and videos geotagged in the month this paper was written[3]. Flickr allows users to provide geolocation information for their pictures either as

[1] http://zonetag.research.yahoo.com

[2] http://tagmaps.research.yahoo.com/



(a)



(b)

Figure 1: Two iconic views of our world. (a) The World at Night, and (b) The Photographed World.

exact or approximate geographical coordinates with the help of a map interface or as geographically relevant keywords. Geotagging can also be performed by using a digital camera or smart phone equipped with a GPS receiving sensor or by using a digital camera that can communicate with a standalone GPS receiver (e.g., through a Bluetooth® link). Photos can also be synchronized with a GPS logging device.

It has become a serious challenge to manage such an overwhelming amount of multimedia data. Currently, commercial search engines and Web albums rely on text annotations associated with images for indexing and retrieval tasks. Richer and more accurate semantic annotation would benefit many applications including image search, sharing, organization, and management [3]. Recognizing the need for semantic annotation, the latest version of the Google™ Picasa™ now enables users to label images in terms of faces, places, and user-specified tags, as shown in Fig. 2.

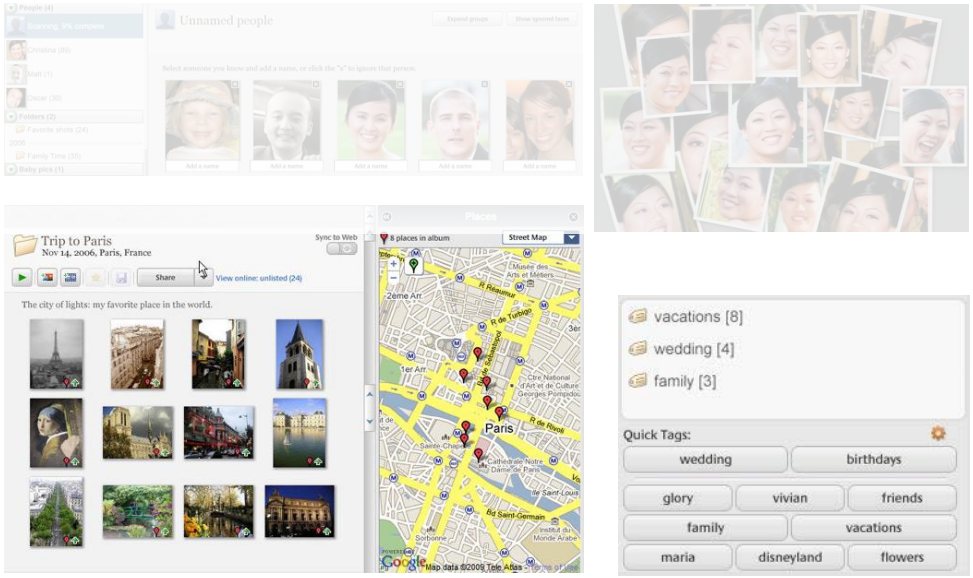


Figure 2: Personal images are often described in terms of people, places, and events (courtesy Google Picasa).

In the last several years, an important trend that has been noticed within multimedia understanding and computer vision domains is the increased emphasis on modeling and using contextual information [23,56,96,100,101]. Contextual modeling often gives an opportunity to derive more leverage from data than pixels alone can provide. A few sources of contextual information are: (i) meta-data captured with pictures or videos [56,94], (ii) relationships between spatio-temporal segments in multimedia data [56], or (iii) patterns in multimedia collections as a whole [11,12,105]. Geographic information has been embraced by multimedia and vision researchers within a contextual modeling framework. However, possibilities of modeling and extracting value from geographical context are many and vary as will be discussed.

Figure 3 shows the structure of our paper schematically. In Section 2, we discuss modalities of geo-information that have been found valuable in multimedia and vision research. Section 3 forms the majority of our discussion and focuses on geotagging driven multimedia applications wherein the roles of geo-modalities are implicitly discussed. We would like readers to note that while Fig. 3 shows a fully connected bi-partite graph between geo-modalities and applications, all links may not appear equally weighted in our discussion. We believe that this is justified for two reasons: (a) geotagging is relatively new in multimedia domain and the geo-modalities that we discuss in Section 2 may not have been fully exploited yet in multimedia applications; (b) all modalities may not be suitable for all applications. Section 4 marks the conclusion of our paper and lays out certain trends and directions for future research.

2. Geo-Information Modalities in Multimedia Research

We have seen that GPS co-ordinates allow the possibility of images to be mapped on the globe (Fig. 1 (b)). If the globe were to be treated as a rectangular co-ordinate system, inference such as assessing the popularity (density of pictures) of picture taking locations or studying location traces in collections can be directly performed with GPS co-ordinates for making inferences about individual images or image collections as a whole. GPS co-ordinates can also be used as keys into other forms of geographic knowledge (from personal, public, specialized, or community sources [27]) that can enable finer understanding of images or collections. In literature, we find a variety of geographic knowledge spanning a wide spectrum from simplistic to complex, direct to implicit, textual to numeric, and individual to aggregate in nature. Here, we attempt to categorize and discuss the geographic knowledge sources or modalities based on how they have been used in the work that we survey in this paper. In Figure 4, we motivate the readers with some graphical visualizations of how geo-modalities are used with respect to certain example applications. We will return to the figure in the following subsections.

2.1. Points of Interest Geographical Databases

Points of interest (POI) or Geographical Information Systems (GIS) databases provide useful geographical knowledge about important landmarks or points of interest. A GIS database typically consists of detailed geographical information about places, such as latitude, longitude, country, population, and other demographics. Geonames is a GIS database available free of charge and

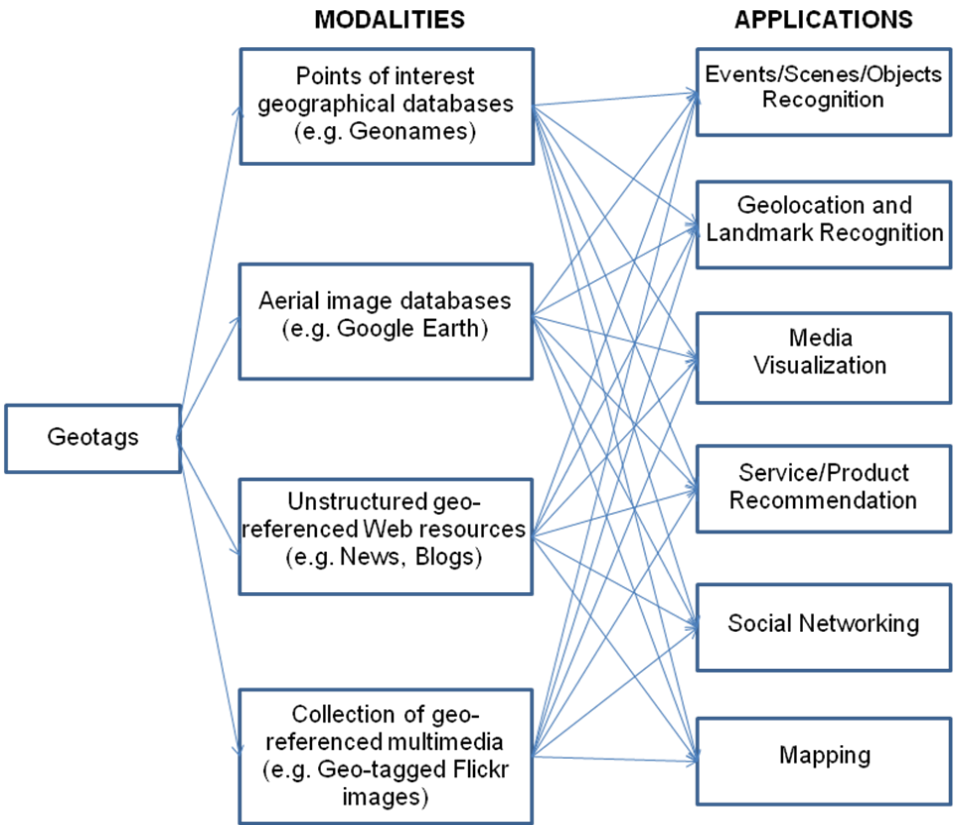


Figure 3: Geo-information modalities and geotagging driven applications in multimedia research.

contains over 8 million geographical names and associated information collected and compiled from multiple sources. The process of querying a POI database given a GPS co-ordinate is known as reverse-geocoding. POIs can be used for providing geo-location based services to travelers [107], performing intelligent actions based on location demographics [52], summarizing the location with nearby points of interest tags to assist image understanding [40] (also depicted in Fig. 4(a)), or identifying persons in images [65].

Place-names and corresponding demographics, being purely textual or numeric entities, can inherently be handled using text or data analysis techniques. At the same time, GIS databases contain fairly reliable data collected professionally with serious human effort. Intuitively, reverse geocoding through entries in a GIS database should be specific enough to identify the location environment and help summarize a location. However, certain limitations to acquiring complete semantic knowledge through this process are: (1) A place is represented as a point (e.g., the central office in a zoo) in the database without any definition of the actual spatial extent; (2) Multiple environments can potentially co-locate in close proximity of each other – such as cemeteries, churches, schools etc; (3) GIS databases may miss details such as tennis courts or parking lots inside schools or parks.

2.2. Aerial Image Databases

Aerial image databases present yet another modality that can be tapped for acquiring and using geographical knowledge. Aerial image databases such as Google Earth and Microsoft's Bird's Eye View have virtually captured the entire globe at varying levels of detail and can reconstruct the terrain at user's will. Such systems combine satellite imagery, aerial photography, and map data to make a virtual, interactive template of the world. Besides providing a wonderful virtual flying experience through the world, Google Earth also allows users to geotag their personal multimedia or view geotagged images and videos taken by others. Aerial images acquired from Google™ Earth can provide useful cues as to the location environment that can complement multimedia understanding systems. Researchers have used machine learning to infer the location environment

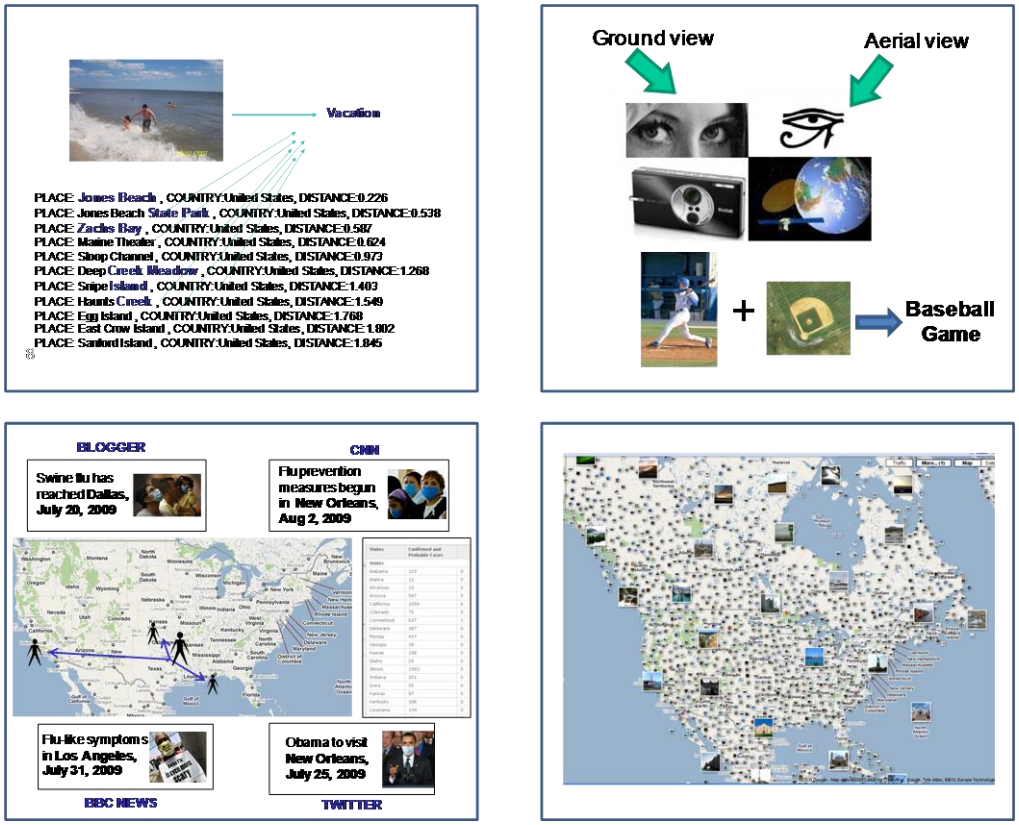


Figure 4: Example applications depicting use of geo-information modalities (a - top left) Points of interest geographical databases, (b - top right) Aerial image databases, (c - bottom left) Unstructured geo-referenced web resources, and (d - bottom right) Collection of geo-referenced multimedia.

from aerial pictures as a key step towards event-activity recognition [57] (Fig. 4(b)). Although not an aerial image database, Google Street View is also a comprehensive multimedia resource from Google that is quite thorough in its coverage of the US and has been used for making viewing directional inference [67].

The main advantage of employing aerial images to recognize picture-taking environments is apparent: free of distraction and influence by the actual event depicted in a ground-photograph (i.e. taken at ground level), which is often an adverse factor for computer vision algorithms. In fact, the aerial image is practically invariant relative to the current event, weather, and participants. However, complete reliance on aerial images can be risky especially for the following reasons: (1) Aerial images that capture location structure and environment often require very precise location information. Yet geotagging can be noisy because of inaccuracies in GPS sensors or human tagging. Errors in GPS co-ordinates can result in a shift in the corresponding aerial image that could at times lead to incorrect prediction; (2) Aerial images capture the environment of the picture taking location (of the camera) rather than the location depicted in the picture. Therefore for scenarios such as shots of far-away objects or scenes, dependence on aerial images can be rather misleading.

2.3. Unstructured Geo-referenced Web Resources

Websites, news, blogs, wikis, and social networks that contain information with geographic bearings can also be useful sources of knowledge for multimedia research. These can provide information of extremely diverse nature (in content and modality) contributed by common people (non-experts). Information can also be current and conversational about live events or major happenings. As an example, the micro-blogging service Twitter now provides a way to geo-tag one's tweets by choice. Generic geotagged information such as tweets can be used to identify, gauge, and study community interests and reactions about hot-topics across space and time. Such analysis can help facilitate recommendation or prioritization of health services based on socio- and spatio-temporal behavior of information with respect to events such as spread of swine-flu [87] (Fig. 4(c)) or regional product popularity and political polling [39].

Information from community sources, though vast and diverse, comes with the following concerns: (1) Because such sources are constructed, contributed, and moderated by masses they are prone to a variety of noise. The verity of knowledge acquired from them cannot be easily established. Moreover, the objectivity of knowledge can also become questionable as there may be political, regional, communal, or personal biases; (2) Information from such sources is multi-modal and unstructured. Making the information fit for multimedia research and applications would typically require a heavy pre-processing overhead.

2.4. Collection of Geo-referenced Multimedia

Collections of geo-referenced images and videos (as opposed to individual pieces of image or video data) have proven very useful in multimedia research. Such collections are becoming increasingly popular and accessible thanks to photo-sharing services such as Flickr and Google Picasaweb that have realized the need to tap into geographical information for search, sharing, and visualization of multimedia data. The interesting prospects that community geo-referenced multimedia databases present are: (i) The possibility of using brute force simplistic search methods for inferencing [31]; (ii) Constructing semantic geographical summaries using visual and meta-data information [26][61]. Given a geographic location, meaningful labels can be obtained from pictures taken in the vicinity of the location [61] (Fig. 4(d)). The sheer volume of data can also help build trustworthy models for multimedia understanding [19]. While personal geotagged collections are typically small, they present the prospect of leveraging the correlations between location and time collectively to assign labels to individual pictures or picture collections at a personal scale [11].

Among the papers that we review in this paper, the current geo-modality is the most abundant. While the human-assigned tags to pictures may be considered largely reliable here, certain pertinent challenges still persist in the form of: (1) Incorrect or alternate spellings associated with certain place-names (e.g., Eifel tower and Eiffel tower are both popular); (2) Inherent polysemy and synonymy of words in the English language.

3. Geotagging Driven Applications in Multimedia and Vision Research

3.1. Semantic Multimedia Understanding – Events, Scenes, and Objects Annotation, Organization & Retrieval

Semantic multimedia understanding is a generic term encompassing a variety of research problems. As is evident from the name itself, understanding pertains to semantic interpretation of primarily the content of an image or video (sometimes audio and text data are also involved). Broadly speaking, semantics in images and video has been the subject of much computer vision research in the last decade. In multimedia research, concepts, scenes, or events define a controlled set of entities into which the pictures or video at hand can be necessarily classified [11,40]. An alternative paradigm, image annotation or tagging allows image labels to be less restrictive, free-form, unstructured, and hence spawning a larger often user-given vocabulary with polysemy and synonymy among words [13]. A related and often co-discussed research issue is that of multimedia retrieval since retrieval engines derive their bearings from the picture content itself [5,6,7]. Concept, event, scene detection or annotation are all essential to understanding, managing, and selectively retrieving images and video from personal collections or large repositories, and hence research in retrieval is heavily tied to understanding of semantics [97].

3.1.1 Event/Scene Understanding

A key form of semantic understanding in the multimedia realm is identification of the type of event that the user has captured, such as a birthday party, a baseball game, a concert, and many others where pictures and videos are commonly taken. Typically, events such as these are recognized by using classifiers that are learned using a set of training images to permit the computation of a decision or a probability that a new image of concern is of a certain known event type. Here we discuss works that employ geographical context in innovative ways for event and scene understanding.

Geometrical properties of traces put together from locations of pictures from collections can encode useful information to describe the event type. In [105], the authors use geographical trace information in association with certain informative compositional visual features to detect event types such as hiking, skiing, party, etc. Clearly, the inherent information that is tapped here is the motion patterns of the photographers and the photographed. Another source of derived knowledge

is geographical databases discussed in Section 2.1. In [53], Liu et al. proposed a personalized location recognition system that identifies semantic locations such as home, office, friends' or relative's home or business location using map traces and reverse geo-coding. Points of interest summaries were used to find associations between places, picture semantics, and location in [40]. The aforementioned work models salient geographical descriptive words or geotags with respect to generic event categories such as vacation, beach, hiking, etc. Aerial images, the knowledge modality described in Section 2.2, have been employed by Luo et al. in [57] for event recognition. The paper illustrates how ground images and the co-located aerial images complement one another for event recognition. Another interesting work [104] acknowledges seasonal and geographic differences for improved scene understanding. The basic motivation lies in the fact that seasonal and geographic information with pictures can help improve region labeling. For example, a white region is more likely to be snow in North Eastern US in the winters whereas it is more likely to be sand/cloud in Florida all year round.

While analyzing an image is an indispensible part of the recognition process, patterns mined from image collections as a whole can often shed additional light on the events captured in them. Earlier we saw the use of image collections for deriving trace information in [105]. In [11], a hierarchical image annotation technique is explored for geotagged image collections. The distinguishing aspects of the work include incorporating (i) the relationship between collection-level annotations (events) and picture level annotations (scenes), and (ii) the time-location constraints, into the modeling and annotation process. In a similar vein, [12] exploits image collection level semantic coherence for labels propagation. An unsupervised engine propagates the high confidence labels (obtained with supervised classification) using visual, geographical, and temporal similarity between pictures in a collection. In [76], visual, textual, and spatial information has been used to perform object and event mining on a large-scale geotagged image collection. The work derives its novelty from linking semantic image clusters to articles in Wikipedia and reaffirming cluster assignments using content in Wikipedia pages.

3.1.2 Multimedia Organization, Annotation, and Retrieval

Semantic organization of multimedia has been one of the priced pursuits of the multimedia research community. While complete multimedia semantics is still an open problem in itself [69], meaningful semantic organization, in its current scope, usually entails grouping entities based on their visual appearance and other attributes. One of the earliest works employing geo-location information for image organization was proposed by Naaman et al. in [62]. The cited paper used spatial and temporal information to produce location and event hierarchies. Similarly, Pigeau et al. employed hierarchical geographical and temporal partitions for image collection management on mobile devices [72]. In a more recent work, Ephstein et al. describe an innovative approach that uses location and orientation information obtained from multiple cameras to discover geographically salient objects [24]. This is followed by a hierarchical organization of scenes or views for efficient browsing.

While research using geotagged images has gathered considerable momentum, geotagged videos are still a relatively rare entity. Indexing, organization, and retrieval of videos captured using wearable cameras, GPS receivers, and gyroscope was introduced in as early as 2002 in [99]. Certain recent papers again focus on geo-referenced videos [5,6,7,46] and employ spatial and temporal video properties for ranking and search. Under the experiments described, the cameras collecting video data are equipped with GPS sensors and 3D digital compasses. The goal is to capture and model the "viewable scenes" or the "fields of view (FOVs)" relative to a location. A user typically interacts with the system by specifying a query region using a point, a line, or some form of a polygon area on a map. The retrieval process involves (i) performing spatio-temporal filtering to get relevant videos, (ii) ranking videos by the overlap of their FOVs with the query region. On a similar note, spatial queries using map interface have been used in [22] for story retrieval.

Image annotation or tagging research has also been increasingly focused around geotagging. Naturally, annotation models require establishment of structure or association among words or tags and hence natural language processing is often necessary. A location dependent pLSA model (LD-pLSA) was proposed for joint modeling of visual and geographical information in [20]. The model attempts to extract two different kinds of latent classes, visual topics and geo-topics (the latter influencing the former), underlying the observed data. The work aims to perform a geo-visually coherent image categorization automatically using the dual topic modeling approach. Annotation of pictures using very large collections of community geo-referenced pictures is vastly becoming popular in multimedia research. In [47], a 1.2 million Flickr image collection is used to build a geo-profile which is then employed to annotate images. The first step involves prediction of geographic co-ordinates of the input image using K-NN approach as in [31] that the user can

Figure 5: The first three scenes are iconic and well-known landmarks (Eiffel tower, Washington Monument, and Sydney Opera House) that are relatively easy to geo-locate. The other images contain cultural features that only allow for coarse geo-location.

choose to refine. The geographic location is then used to suggest likely tags in the identified location as learned from the image collection. In rhyme with the above work, [61] presents Spirit-tagger tool that mines tags from geographic and visual information. The sheer volume and diversity of geotagged images on the Web have been used to explore and exploit deeper correlations between visual features and their annotations using Logistic Canonical Correlation Regression (LCCR) in [13].

3.1.3 Social and Cultural Semantics in Multimedia

Location context has also been shown to assist identification of people in pictures. One of the earliest attempts toward using geographical context for face recognition was made in [21]. The cited work geographically clusters images with GPS and uses visual, spatial and temporal cues for prediction. A similar approach is also explored in [63] where co-occurrence and re-occurrence patterns of people with respect to locations and events have been modeled for resolving people identities in consumer collections. A content and context based photo management system MediAssist [64,65] employs contextual information such as clothing and geographical information for person identification. However, the previous work does not handle multiple people in an image, a problem that has been explored in detail in [25]. Yanai et al. have given geotagged image mining an interesting twist in [102,103] by discovering cultural semantics across different geographical regions. The authors visually depict and discuss regional, visual, and semantic variations of words like cake, noodles, etc.

3.2. Geolocation and Landmark Recognition

It is usually easy for a person to identify the location of an image of a well-known landmark such as the Eiffel tower, the Washington Monument, or the Sydney Opera House [75]. Even when an image does not contain a well known landmark, there is often enough information in the image for a person to identify the region of the world; or at least to take an educated guess at the location of the image. For example, printed text in a scene often conveys the local language, which is strongly related to position on the globe. Or, the building materials, architecture, and natural features may supply clues related to the scene location. In the application category of *geolocation recognition*, the goal is to determine the (unknown) location of an image, video, or series of images. The problem can be subdivided into the finer categories of *landmark recognition* and *non-landmark location recognition*. In either case, collections of geotagged images are employed to help infer the location of unknown images. Further features that are exploited include text tags, and sequential image captures.

Because landmarks are just that (i.e. stationary and somewhat unique objects), recognition of landmarks is equivalent to determining the location of a test image. However, landmark recognition is a subset of the work directed toward determining the location of an image. Typically, these algorithms leverage collections geo-located images for training or matching. Further, landmark recognition method generally attempt to directly match feature points or structures from an unknown image to images having known geolocations.

Of course, estimating camera geolocation is closely related to the fields of camera calibration and bundle adjustment, for example [28], where image features that are visible in two or more images of a scene are used to recover both the coordinates of the image features in an external 3D coordinate frame and recover the camera parameters (also including the location of each cameras in the 3D coordinate frame.) This approach, directed toward 3D reconstruction, is explored, for example, in [80][98] and more recently [89]. When this 3D coordinate frame is placed within the context of the coordinate system of the Earth, geolocation recognition has been accomplished.

It should also be noted that many commercially available cameras, especially cameras in cellular phones, can determine location of the user by hardware (GPS hardware, or cellular tower

proximity). The GPS location can be automatically embedded into the image header. In some cases (e.g. with the iPhone), the compass direction of the camera can also be recorded. On one hand, these devices provide a method for tagging images with location without analysis of image data. However, determining the location of an image *from the image data* is still an important research question for several reasons. First, many images have been or will be captured without this hardware-based technology. Second, it is important to know what location information can be extracted from the image itself for either ensuring privacy or security.

3.2.1 Landmark recognition with feature point matching

Early work on determining geolocation from matching regions across multiple images was performed for the "Where am I?" contest at ICCV 2005 [91], where a series of geotagged images were presented to an algorithm for training. The task was to determine the GPS coordinates of each test image, all of which contained at least some overlap with a training image. The winning approach [106] (and indeed also the runners up) implemented a framework where interest points are found in all images. The interest points (such as SIFT [55], SURF [9], or MSER [60]) from a test image are matched to interest points in one or more training set images. In [106], the location estimate is further refined using a triangulation procedure based on planar homographies to the top image matches.

Schindler et al. [82] extends this work to investigate the use of large vocabulary trees for finding the location of query image by matching to a training set of 30,000 images. A vocabulary tree is used to organize the large number of detected interest points by hierarchically clustering the data points at each level of a tree. To efficiently find matches, at each level the tree's nodes are compared to the SIFT features of the query image to find the closest matches. In this work, the tree has a large number of nodes (10^6) and during construction, information gain is used to select visual words that are informative about a particular location.

Interest point-based approaches are extended to large scale recognition of many tourist landmarks in [51,110]. In [110], blogs (Wikitravel) and geotagged images from the Internet are mined to establish a list of 5312 tourist landmarks that are recognized with nearest neighbor. In [51], a set of 500 landmarks are discovered from Flickr images, and recognition is performed with a structured SVM.

Landmark recognition can also incorporate the 3D information related to the landmark itself. In [50], a landmark recognition algorithm is described where recognition is performed by first finding a set of candidate matches (to a set of automatically discovered iconic images that represent each landmark) to a test image using either interest points or GIST [66]. Next, each candidate receives a score (the number of inlying interest points between a two-view transformation) that is based on a geometric verification between the candidate and the test image. Then, the top scoring image is used to determine the landmark classification of the test image. Although tested on only a small number of landmarks, the improvement produced by the geometric verification is startling, and is present for either feature type.

One difficulty of using interest point matching in recognition is that in urban environments, structures tend to repeat many times on one or multiple surfaces. This results from the repetitive nature of building facades (e.g. the same window often covers an entire façade). Therefore, matching a single interest point from a query image does not necessarily pinpoint the location of the camera. In [83], the repeating nature of building facades is exploited by detecting the façade lattice and matching the lattice rather than individual points. This allows for the further benefit of recovering camera pose, so location, compass heading and tilt are all recovered. In other scenarios, the location of a single image or set of images can be estimated with surprising accuracy.

3.2.2 Non-landmark location recognition

Several interesting approaches have been devised to perform *non-landmark location recognition*. In this problem, a training dataset cannot be relied on to contain an exact match to a particular image. In [31], a method is presented for finding a probability distribution on the globe for the location of an unknown image. The idea is to use a data-driven scene matching approach. An image is summarized with typical scene descriptors (e.g. color histograms, GIST [66], texton histograms) and compared to a dataset of over 6 million GPS-tagged images. The approach is able to estimate the locations of images of non-distinct European alleys, tropical beaches, and African savannah scenes not by finding exact feature point correspondence, but rather by summarizing the likelihood that such an image could be captured in a particular part of the world. In essence, the training dataset is used to represent the general appearance (e.g. the color, the amount of vegetation, the type of architecture) of different locations. Surprisingly, the estimated location is correct to within 200 km about 16% of the time. This work was extended in [26,41,84] by also

considering text tags associated with the images to improve performance. In another extension [42], a sequence of images from the same photographer is used along with a model that represents traveling from one location to the next in a given time interval. By jointly geo-locating the series of images, performance is further improved. Note that a similar result was achieved for landmark recognition in [51].

3.2.3 Location recognition from stationary cameras

A creative approach was used to solve the geolocation estimation problem by using an image series from a stationary web camera [35]. Rather than relying on the scene content itself, their method exploits knowledge of solar geometry to correlate the patterns of daylight an night in the images from a web camera to determine the location of the camera on the globe. Most cameras are geo-located to within 50 miles of the actual location. A variant of this approach is described in [92], where the sun's position is recovered by observing its photometric effect (i.e. color changes) and on the scene. When the image capture time is known, this provides both geo-location and geo-orientation without requiring the sun to be in the camera's field of view.

3.3. Media Visualization

With the availability of associated geotags, photos and other media can be visualized in a whole new dimension by location. This can be accomplished in a number of ways, including visual summary of media collections, visual summary of landmarks, visualization of travel trajectories and routes, visualization of tourism and life patterns, as well as more sophisticated visualization in terms of viewing directions and 3D experiences. While most of the visualization is geared towards tourism, it can also benefit visual recognition and other applications.

3.3.1 Visual summary of collections

Torniai et al. [93] propose an early system to provide a new browsing experience of photo collections based on location and image header metadata. For a large collection of geo-referenced photographs, Jaffe et al. [36] develop a framework for automatically selecting a summary set of photos, because such large collections would require making summaries in order to be rendered accessible. Their summary algorithm is based on spatial patterns in photo sets, textual-topical patterns, and user (photographer) identity cues. In addition, they present a modified version called Tag Maps, which serves as a basis for a new map-based visualization of large collections of geo-referenced photos. In essence, the Tag Maps visualize the data by placing highly representative textual tags on relevant map locations, providing an idea of the location distribution of important concepts embodied in the collection.

3.3.2 Visual summary of landmarks

In a broader context, a series of work has been reported on generating visual summaries of landmarks using geotagged photos on the web. In [44,45], Kennedy et al. use the bag of visual words approach to generate diverse and representative images, or a visual summary, of a location based on geotagged community images. More recently, Chen et al. [17] further generate a tourist map by automatically identifying popular points of interest (POIs) from community photo collections, and rendering the representative landmark icons based on importance. At a larger scale, Crandall et al. [19] use a dataset of about 35 million images collected from Flickr. Their system combines content analysis based on text tags and image data with structural analysis based on geospatial data. They exploit the interplay between this structure and the content, using classification methods to predict landmark locations from visual, textual and temporal features of the photos. They were able to discover and display various interesting properties about popular cities and landmarks at a global scale. Ji et al. [37] use an alternative resource of blogs to mine popular city landmarks for personalized tourist suggestions. The main idea is a graph modeling framework to discover city landmarks by mining blog photo correlations with community supervision. More interestingly, they also investigate how city popularities, user locations, and sequential events (e.g. Olympic Games) influence their Landmark discovery results and the tourist suggestions.

3.3.3 Visualization with camera viewing directions

As discussed in Section 3.1, recent research results have shown that the additional GPS information helps visual recognition for geotagged photos by providing location context. However, the current GPS data only identifies the camera location, leaving the viewing direction uncertain. Even with a digital compass, the recorded camera direction may still be inaccurate (Nokia Challenge: http://comminfo.rutgers.edu/conferences/mmchallenge/2010/02/10/nokia-

challenge/). For a given geo-location, many photos with view directions not pointing to the desired regions are typically returned through location-based search. To address this problem, a method is presented in [59] to estimate the viewing directions of photos and furthermore obtain the camera pose in 2D referenced on the Google Maps using the geographic metadata of photos. The main idea is to first generate subsets of images from a large number of photos taken near a place, and then reconstruct the scene expressed by those subsets using a normalized 8-point algorithm to estimate the camera rotation and translation. In particular, this system extracts SIFT features of all photos in the database, indexing them using the KD-tree algorithm. Given a query example of an underlying scene, a set of matched photos of a place are found from the database using SIFT Flow. Finally, after registering the photos by viewing directions, the photos with view directions pointing to the user-indicated region are returned by the system called ViewFocus [58]. More recently, to produce more precise location information, including the viewing direction, for a *single* geotagged photo, Park et al. utilize both Google Street View and Google Earth satellite images [67]. Their system is two-pronged: 1) visual matching between a user photo and any available street views in the vicinity determines the viewing direction, and 2) when only an overhead satellite view is available, near-orthogonal view matching between the user photo and corresponding overhead satellite imagery estimates the viewing direction. A related camera pose estimating system based on 2D building outline maps is described in [15], as well as another system that relies on vanishing points [48].

3.3.4 Visualization of travel trajectories and routes

GPS data can help visualize locations related to media. Moreover, the trajectories or traces formed by connecting a sequence of GPS locations can help visualize people's movements (sometimes as they capture images and videos) [16, 70], and are further useful for inferring high-level activities. To remove the uncertainty or noise in the GPS trajectories due to GPS errors, Pelekis et al. [71] developed a method for clustering trajectories of moving objects and then computing the centroid trajectory of a group of movements. In [52], a typical trace consists of approximately one GPS reading per second. A GPS trace is segmented to generate a discrete sequence of activity nodes, each of which represents a set of consecutive GPS readings that are within a certain area. Subsequently, one can extract a person's activities and significant places from traces of GPS data. Also using GPS log data, Zheng et al. [109] perform data mining to discover interesting locations and travel sequences from GPS trajectories of many users. In a similar vein but using only sparse (as opposed to continuous) GPS traces formed from the GPS locations of a relatively small number of geotagged photos, Yuan et al. [105] show that even such GPS traces are useful for characterizing the underlying movement patterns of various event and activity types. They further derive informative features from a discrete sequence of GPS coordinates and a bag of visual words, both based on the entire collection as opposed to individual photos, for improved event recognition.

3.3.5 Visualization for Photo Tourism

In a series of ground breaking work [85,86,88,89,90], Snavenly et al. developed a system commonly referred to as Photo Tourism. It is a system for browsing large collections of photographs in 3D. Their approach takes as input large collections of images from either personal photo collections or Internet photo sharing sites, and automatically computes each photo's viewpoint and a sparse 3D model of the scene using robust SIFT features. Upon scene reconstruction, their photo explorer interface enables a viewer to interactively move about the 3D space by seamlessly transitioning between photographs, based on user control. More recently [1], they have scaled up the algorithms behind Photo Tourism using a parallel distributed matching system to work on an entire city with a million or more images. Such large scale data collection represents an increasingly complete photographic record of the city, capturing every popular site, facade, interior, fountain, sculpture, painting, cafe, and so on. Therefore it becomes possible to richly (and virtually) reconstruct, explore and study the three dimensional shape of the city. In a related work [43], the same research group also investigates ways to align 3D Point Clouds to overhead satellite images. At a personal level, to help people relive their travel experience they had recorded in photos, Hsieh et al. [34] present a system called Photo Navigator to enhance photo browsing experience by creating a new browsing style with a realistic feel to users as being into the scenes and taking a trip back in time to revisit the place. This system is characterized by two main features. First, it reveals the spatial relations among photos and offers a strong sense of space by allowing users to fly into the scenes. Second, it is fully automatic and makes easy for users to utilize the 3D technologies that are traditionally complex to manipulate.

3.4. Service/Product Recommendation

Geo-tagged media such as image, video and blogs often reveal important information for many location-based service or products. For instance, people can select their vacation destinations based on web images and travelogues. They can also be prompted of local events or attractions based on their current locations and interests. Research works in this direction share a general goal of mining the user generated contents for geo-services, which needs to overcome many challenges such as noisy content, location ambiguity, and semantic gap. The subjects of the research work in this area can be further categorized into four groups: (1) real-time recommendation delivery, recommendation inference via (2) geo-tagged images (3) travelogues and (4) GPS trajectories, which is discussed in detail in the following subsections.

3.4.1 Real-time recommendation delivery

Early works in this area focus on the software infrastructure of integrating location sensing and recommendation delivery, where the recommendation is assumed to be stored in or automatically generated from an existing database. For example, Chen et al. study the integration of location-aware services from a software system perspective in [18]. They design a location operation reference model (LORE) as a middleware to support many location-related services and operations such as location data capture and fusion, location tracking and notification (as in [54]). Interestingly, they also spend considerate effort on implementing control mechanism for privacy, security and management issue of client location data, which has not been widely studied by other work in the literature. In another work, Hinze et al. make a notable contribution of considering user profiles and query histories in order to extract personalized tourism information [33]. They also note that privacy issues in leveraging user profiles pose challenges and should receive significant attention.

3.4.2 Travelogue mining

Many online communities and blogs contain travelogues that provide first-hand user comments on tourism attractions, activities and even costs. Pioneering research work in this area has been conducted by a group of researchers in Microsoft Research Asia. They demonstrate a travel recommendation system called “TravelScope” in [29]. It collects and analyses over 20000 travelogues on the web to first identify cultural and natural attractions. For each attraction, it can further extract the representative views that include associated popular tags and photos. When user selects/clicks on a picture or tag, the system also displays the original travelogues that mentioned them to provide more detailed information. The authors also introduce a method to detect the attractions. It uses a generative framework to model the latent topics that governs the distribution of words in travelogues associated with different locations. Different from well-known techniques such as PLSA, the authors divide the topics into two categories: global and local. The paper defines the global concepts, such as taxi and airport, as shared by most locations and the local ones such as beach as specific to only a subset of locations. To extract local topics, mutual information is further adopted to model the correlation between local topics and locations. Finally, the tags generated by the local topics along with the location names are used to retrieve the representative images for different locations. More recently, in [30], a more comprehensive model is designed to model the local and global topics depicted in travelogues. Instead of using bag-of-words as in the previous work, it assumes a block of consecutive words, e.g. words in paragraphs and sentences, share the same location and topic assignment. Fitting such a model allows the estimation of local topic distribution for each location and query words. Therefore, the similarity between location and query words can be evaluated in the local topic space that is relevant to attraction recommendation tasks. In addition, new methods for discovering representative words are designed using the learned distribution of location and words.

3.4.3 Geo-tagged image mining

Thanks to advocating by commercial web services such as Yahoo Flickr and Google Picasa, geo-tagged images are arguably the most popular geo-tagged multimedia format and attract a lot of attention from the research community.

Recommending tourism locations from large-scale geo-tagged image database is investigated by Cao et al. [14]. They match user-provided images or tags with the representative ones from the image database, which naturally requires efficient clustering of the images. To achieve that goal, a new flat kernel is designed to measure the similarity of 2-D coordinate vectors. The representative images are selected from individual clusters using affinity propagation. Human evaluation of the recommendation accuracy demonstrates the feasibility of the proposed method. To mine representative landmarks and produce personalized tourism recommendations, Ji et al. [37] model

the relationship of scene/landmark, scene and authorship as a graph and adopt two popular link analysis methods, PageRank and HITS. They specifically design a method to extract tourism information from millions of travel blogs on the web. It is worth mentioning that on these blogs the photos are not explicitly geo-tagged as in photo-sharing websites. However, the relationship between photos and locations can still be inferred by matching image titles and surrounding text with a gazetteer. For a certain landmark, the representative photo is selected using a PhotoRank method, which extends the well-known PageRank method by weighting the visual and text words based on their importance. To extract the landmarks in cities, the authors integrate authorship with photo and scene similarity in a HITS model.

Although most of the geo-related research focuses on outdoor images, indoor geo-tagged images may also carry important location information. Paucher et al. [68] leverage geo-tagged indoor images along with the accelerometer and magnetometer on cell phone to realize augmented reality on mobile devices. They use GPS to locate the rough location of a user in an indoor environment where sample images have been extracted. By matching the user captured images with the exemplars, the pose that is initially estimated by the accelerometer and magnetometer can be significantly refined. It is worth mentioning that their algorithm requires minimal computational power and can run on the mobile phones.

Different from the above mentioned works that analyze the visual content of geo-tagged images for tourist suggestion, Popescu et al. [73,74] leverage only the context information. It aims at mining the geo-tagged images to find visiting time needed for popular attractions. Comparing with the visit time recorded by real users, the inferred time is considered relatively reliable. Similarly, Rattenbury et al. show that analyzing only the flickr tags can discover two important semantics: "event" and "place" [77,78]. They define the event tags as having significant temporal patterns and place tags as having significant spatial patterns. After analyzing the location and time metadata available in geo-tagged images on Flickr, the authors were able to detect the temporal and spatial patterns and categorizing the relevant tags with reasonable accuracy.

3.4.4 Geo-trajectory mining

With the increasing popularity of geo-enabled devices, it is possible to record user's travel experience with precise GPS coordinates. It is desirable to recommend suitable locations and activities based on user interest and the wisdom of the community. Zheng et al. [109] pioneer the research on analyzing tourists' GPS trajectories and finding interesting locations and visiting sequences. Their work starts with clustering the locations in a hierarchical order so that tourism recommendation can be conducted at different scales. First, the users and locations are connected by the visit activities in the trajectories. Then, they propose to use HITS-based method to calculate the hub and authority scores for users and locations, respectively. Combining these two scores can further rank the visit sequences in the GPS trajectories. Zheng et al. [107] extend this work to tackle the problem of suggest popular locations/query for the activities/locations of user query. Interest regions are also clustered from the user trajectories and the corresponding activities can be extracted from user comments. However, the big problem is that such correspondence is sparse. The authors propose to first extract more information about the interest regions and activities correlations by mining related web-pages. Then, the enriched information is fused together by factorizing the correspondence matrix into the product of two low-rank matrixes. In another extension, Zheng et al. [108] introduce a social networking service, called GeoLife, through understanding trajectories, locations and users, as well as the correlation between users and locations. GeoLife allows sharing life experiences based on GPS trajectories; popular travel recommendations based on travel sequences and travel experts in a given region, and finally personalized friend and location recommendation.

In many applications, the geo-related information is stored in the form of GPS coordinates or precise street address. However, the semantic meaning of the location to a user, e.g. home, office, interested restaurant, is still unclear. Liu et al. [53] propose a method to extract such information from GPS trajectories and extensive metadata that include user calendar, address book, phone call list and user feedback. After analyzing the outdoor GPS trajectories of a user, it first finds the important locations where user stays for longer than a pre-defined threshold. To extract the semantic meaning of these locations, reverse geo-coding is applied to obtain a set of business names for the GPS coordinates or street addresses. A rule-based matching method is further developed to predict the semantic location type based on user profile information such as calendar and phone call history. It is clear that research work on geographical recommendation prospered in the past decades. It is reasonable to expect it will attract even more attention due to the increasing popularity of GPS-enabled phones.

3.5. Social Networking Applications

Recent years witness an explosive growth of social network services. Popular websites such as Facebook and Twitter attract millions of user. The user profiles often contain significant geographic information that can be leveraged to understand their social activities.

Sakaki et al. [79] propose to mine user tweets and discover the time and location information related to certain event such as earthquake and typhoon. They model the users as social sensors and their tweets related to the interested events as noisy responses. Two methods based on Karman filter and particle filter produces fairly reasonable estimation of the location of earthquakes and trajectory of typhoons. Backstrom et al. [8] investigate the relationship of social connections and geography. They find that friends who live close to each other are more likely to have more interactions in online social network communities. Based on that observation, they propose a probabilistic method to estimate the location of a user based on the public location information from his/her friends. Further study indicates that iterative updates of the users' locations produce more prediction accuracy.

Geography information of users' residency and the multimedia content that they published online can jointly discover many events in spatial and temporal domain. Singh et al. [87] consider user's twitter and Flickr uploads as responses from "human sensors" across the globe. By aggregating such responses to certain event/query, they obtain visualization of event awareness called "social pixels". Systematic query algebra is further developed to efficiently mining the social pixels without understanding every details of the aggregation process. Experiments suggest it can effectively visualize the propagation of social events, such as swine flu, in the social world.

3.6. Mapping Applications

The history of using images for gathering data about our surroundings is indeed a rich one. In the 1850s, Gaspar Felix Tourachon (pseudoname Félix Nadar) of France was credited with being the first person to take aerial photographs, and was granted a patent teaching aerial photography. Shortly after cameras became airborne, people began to use cameras for the purpose of mapping. Early use of aerial cameras included military tactics where military operations were planned based on the positions of enemy troops. Technological developments in both flight and photography allowed for steady advances in aerial photography and mapping.

For the purposes of this survey, we are interested in approaches where images are used in an automated fashion to produce maps that indicate the position of either manmade (e.g. roads, buildings, cities) or natural (lakes, rivers, mountains) features. In a broad sense, these applications show "what is where" from images. In order for distances on an image to correspond with actual distances in the world, the image must be rectified so that there is no perspective distortion. The image must correspond to an orthographic perspective of the earth's surface. Images with this quality are called orthophotos, and are produced either with specialized. In many cases, the images capture information in spectral ranges beyond the human visual system sensitivity. Algorithms have been developed to automatically map roads, buildings, parking lots, and other natural features from airborne or satellite images [10][38][95][32]. There is a huge volume of work in this area, and the referenced works should be considered as representative of the field.

Such mapping applications can also employ images captured at ground level. For example, [20] shows that many geo-located images can be clustered by considering both spatial proximity and image similarity to produce segments on a map that look something like political states or countries. In recent work, [49] show that collections of geotagged images from Flickr can be used to produce land usage maps. For example, if a specific location contains images with many green pixels, then it is likely to be an undeveloped area.

4. Conclusions and Future Research Directions

In this paper, we have surveyed over 100 geo-tagging related research papers within the context of multimedia and computer vision. Our discussions encompassed major modalities of geographical information, as well as major geotagging driven applications.

We now summarize several marked trends across the field with future research directions. We categorized the sources of geographic information coarsely as point-of-interest databases, aerial and satellite images, geotagged images, and general geo-referenced multimedia. Certainly, with the availability of the Internet, GPS devices, and smart phones, the proliferation and availability of geotagged media will continue to expand. The applications of geographic information in conjunction with multimedia are numerous, but can be roughly categorized as location recognition, object or event recognition, visualization, recommendation, social networking, and mapping. More generally, many of the applications are directed at helping a computer understand one or more

images, based on known relationships that record which objects are likely to be where in the world. Other applications consolidate a large-scale dataset of geo-tagged information to produce maps that indicate where things are in the world. We expect future geotagging-driven research and applications to develop in several directions, including dealing with large-scale data, fusion of multi-modality information, links to social media, as well as 3D, augmented and immersive experiences.

5. References

1. Agarwal S, Snavely N, Simon I, Seitz S M, Szeliski R (2009) Building Rome in a day. In Proceedings of ICCV.
2. Ahlers D, Boll S (2008) Oh Web image, where art thou? In Proceedings of MMM.
3. Ames M, Naaman M (2007) Why we tag: Motivations for annotation in mobile and online media. In Proceedings of SIGCHI Conference on Human Factors in Computing Systems.
4. Amitay E, Har'El N, Sivan R, Soffer A (2004) Web-a-where: Geotagging web content. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.
5. Arslan S, Zhang L, Kim S H, He M, Zimmermann R (2009) GRVS: A georeferenced video search engine. In Proceedings of ACM Multimedia.
6. Arslan S, Kim S H, He M, Zimmermann R (2010) Relevance ranking in georeferenced video search. Multimedia Systems 16(2):105-125.
7. Arslan S, Zimmermann R, Kim S H (2008) Viewable scene modeling for geospatial video search. In Proceedings of ACM Multimedia.
8. Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of WWW.
9. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In Proceedings of ECCV.
10. Benz U, Hofmann P, Willhauck G, Lingenfelder I, Heynen M (2004) Multiresolution object oriented fuzzy analysis of remote sensing data for GIS information. ISPRS Journal of Photogrammetry and Remote Sensing, 58, 239-258.
11. Cao L, Luo J, Kautz H, Huang T (2008) Annotating collections of geotagged photos using hierarchical event and scene models. In Proceedings of IEEE CVPR.
12. Cao L, Luo J, Huang T S (2008) Annotating photo collections by label propagation according to multiple proximity cues. In Proceedings of ACM Multimedia.
13. Cao L, Yu J, Luo J, Huang T S (2009) Enhancing semantic and geographic annotation of Web images via logistic canonical correlation regression. In Proceedings of ACM Multimedia.
14. Cao L, Luo J, Gallagher A, Jin X, Han J, Huang T S (2010) A worldwide tourism recommendation system based on geotagged web photos. In Proceedings of ICASSP.
15. Cham T J, Ciptadi A, Tan W C, Pham M T, Chia L T(2010) Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map. In Proceedings of IEEE CVPR 2010.
16. Chen L, Ozsu M T, Oria V (2005) Robust and fast similarity search for moving object trajectories. In Proceedings of ACM SIGMOD.
17. Chen W-C, Battestini A, Gelfand N, Setlur V (2009) Visual summaries of popular landmarks from community photo collections. In Proceedings of ACM Multimedia.
18. Chen Y, Chen X Y, Rao F Y, Yu X L, Li Y, Liu D (2004) LORE: An infrastructure to support location-aware services. IBM Journal Of Research and Development 48(5/6):601–616.
19. Crandall D, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In Proceedings of WWW.
20. Cristani M, Perina A, Castellani U, Murino V (2008) Geo-located image analysis using latent representations. In Proceedings of IEEE CVPR.
21. Davis M, Smith M, Canny D, Good N, King S, Jankiraman R (2005) Toward context-aware face recognition. In Proceedings of ACM Multimedia.
22. De Silva G C, Aizawa K (2009) Retrieving multimedia travel stories using location data and spatial queries. In Proceedings of ACM Multimedia.

23. Divvala S, Hoiem D, Hays J, Efros A, Hebert M (2009) An empirical study of context in object detection. In Proceedings of IEEE CVPR.
24. Epshtein B, Ofek E, Wexler Y, Zhang P (2007) Hierarchical photo organization using geo-relevance. In Proceedings of 15th ACM Intl. Symposium on Advances in Geographic Information Systems.
25. Gallagher A (2009) A framework for using context to understanding images of people. Ph. D. Thesis.
26. Gallagher A, Joshi D, Yu J, Luo J (2009) Geo-location inference from image content and user tags. In Proceedings of the IEEE Workshop on Internet Vision (with CVPR).
27. Goodchild M F (2007) Citizens as sensors: The world of volunteered geography. GeoJournal 69(4):211-221.
28. Hartley R, Zisserman A (2004) Multiple view geometry in computer vision. Cambridge University Press.
29. Hao Q, Cai R, Yang J -M, Xiao R, Liu L, Wang S, Zhang L (2009) Travelscope: standing on the shoulders of dedicated travelers. In Proceedings of ACM Multimedia.
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 Proceedings of WWW.
31. Hays J, Efros A (2008) IM2GPS: Estimating geographic information from a single image. In Proceedings of IEEE CVPR.
32. Hinz S, Baumgartner A (2003) Automatic extraction of urban road networks from multi-view aerial imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 83-98.
33. Hinze A, Voisard A (2003) Location and time-based information delivery in tourism, Advances in spatial and temporal databases, Lecture Notes in Computer Science, 2750: 489-507.
34. Hsieh C-C, Cheng W-H, Chang C-H, Chuang Y-Y, Wu J-L (2008) Photo navigator. In Proceedings of ACM Multimedia.
35. Jacobs N, Satkin S, Roman N, Speyer R, Pless R (2007) Geolocating static cameras. In Proceedings of IEEE ICCV.
36. Jaffe A, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. In Proceedings of ACM Multimedia Information Retrieval (MIR) Workshop.
37. Ji R, Xie X, Yao H, Ma W –Y (2009) Mining city landmarks from blogs by graph modeling. In Proceedings of ACM Multimedia.
38. Jin X, Davis D H (2005) An integrated system for automatic road mapping from high-resolution multi-spectral satellite imagery by information fusion. Information Fusion, 6 (4), 257-273.
39. Jin X, Gallagher A, Cao L, Luo J, Han J (2010) The wisdom of social multimedia: using Flickr for prediction and forecast. In Proceedings of ACM Multimedia.
40. Joshi D, Luo J (2008) Inferring generic activities and events from image content and bags of geo-tags. In Proceedings of ACM CIVR.
41. Joshi D, Gallagher A, Yu J, Luo J (2010) Exploring user image tags for geo-location inference. In Proceedings of IEEE ICASSP.
42. Kalogerakis E, Vesselova O, Hays J, Efros A, Hertzmann A (2009) Image sequence geolocation with human travel priors. In Proceedings of IEEE ICCV.
43. Kaminsky R, Snavely N, Seitz S M, Szeliski R (2009) Alignment of 3D point clouds to overhead images. In Proceedings of the IEEE Workshop on Internet Vision (with CVPR).
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 ACM Multimedia.
45. Kennedy L, Naaman M (2008) Generating diverse and representative image search results for landmarks. In Proceedings of WWW.

46. Kim S H, Arslan S, Yu B, Zimmermann R (2010) Vector model in support of versatile georeferenced video search. In Proceedings of ACM Multimedia Systems Conference.
47. Kleban J, Moxley E, Xu J, Manjunath B S (2009) Global annotation on georeferenced photographs. In Proceedings of ACM CIVR.
48. Kosecka J, Zhang W (2002) Video compass, In Proceedings of European Conference on Computer Vision (ECCV).
49. Leung D, Newsame S (2010) Proximate sensing: Inferring what-is-where from georeferenced photo collections. In Proceedings of IEEE CVPR.
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 ECCV.
51. Li Y, Crandall D, Huttenlocher D (2009) Landmark classification in large-scale image collections. In Proceedings of ICCV.
52. Liao L, Fox D, and Kautz, H (2007) Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. International Journal of Robotics Research.
53. Liu L, Wolfson O, Yin H (2006) Extracting semantic location from outdoor positioning systems, In Proceedings of the IEEE International Conference on Mobile Data Management.
54. Lothe P, Bourgeois S, Royer E, Dhome M, Naudet-Collette S (2010) Real-time vehicle global localization with a single camera in dense urban areas: exploitation of coarse 3D city models. In Proceedings of CVPR.
55. Lowe D (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision.
56. Luo J, Boutell M, Brown C (2006) Pictures are not taken in a vacuum: An overview of exploiting context for semantic scene content understanding. IEEE Signal Processing Magazine 23(2):101–114.
57. Luo J, Yu J, Joshi D, Hao W (2008) Event recognition: viewing the world with a third eye. In Proceedings of ACM Multimedia.
58. Luo Z, Li H, Tang J, Hong R, Chua T –S (2009) ViewFocus: Explore places of interests on Google maps using photos with view direction filtering. In Proceedings of ACM Multimedia.
59. Luo Z, Li H, Tang J, Hong R, Chua T –S (2010) Estimating poses of world's photos with geographic metadata. In Proceedings of MMM.
60. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22 (10).
61. Moxley E, Kleban J, Manjunath B S (2008) SpiritTagger: A geo-aware tag suggestion tool mined from Flickr. In Proceedings of ACM Multimedia Information Retrieval (MIR).
62. Naaman M, Song Y -J, Paepcke A, and Garcia-Molina H (2004) Automatic organization for digital photographs with geographic coordinates. In Proceedings of ACM/IEEE-CS Joint Conference on Digital Libraries.
63. Naaman M, Yeh R B, Garcia-Molina H, Paepcke A (2005) Leveraging context to resolve identity in photo albums. In Proceedings of ACM/IEEE-CS Joint Conference on Digital libraries.
64. O'Hare N (2007) Semi-automatic person-annotation in context-aware personal photo-collections. Ph. D. thesis.
65. O'Hare N, Smeaton A (2009) Context-aware person identification in personal photo collections. IEEE Transactions on Multimedia.
66. Oliva A., Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision, 42, 145–175.
67. Park M., Luo J., Collins R. and Liu Y. (2010) Beyond GPS: Determining viewing direction of a geotagged image, n Proceedings of ACM Multimedia.
68. Paucher R, Turk M (2010) Location-based augmented reality on mobile phones. In Proceedings of IEEE CVPR.
69. Pavlidis T (2009) Why meaningful automatic tagging of images is very hard. In Proceedings of IEEE ICME.
70. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu M (2004) Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Trans. Knowl. Data Eng. 16(11):1424-1440.

71. Pelekis N, Kopanakis I, Kotsifakos E E, Frentzos E, Theodoridis Y (2009) Clustering trajectories of moving objects in an uncertain world. In Proceedings of ICDM.
72. Pigeau A, Gelgon M (2005) Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In Proceedings of ACM Multimedia.
73. Popescu A, Grefenstette G (2009) Deducing trip related information from Flickr. In Proceedings of WWW.
74. Popescu A, Grefenstette G, Moëllic P-A (2009) Mining tourist information from user-supplied collections. In Proceedings of CIKM.
75. Popescu A, Moëllic P-A (2009) MonuAnno: Automatic annotation of georeferenced landmarks images. In Proceedings of ACM CIVR.
76. Quack T, Leibe B, Van Gool L (2008) World-scale mining of objects and events from community photo collections. In Proceedings of CIVR.
77. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from Flickr tags. In Proceedings of SIGIR.
78. Rattenbury T, Naaman M (2009) Methods for extracting place semantics from Flickr tags. ACM Transactions on the Web 3(1):1-30.
79. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of International Conference on WWW.
80. Schaffalitzky F, Zisserman A (2002) Multi-view matching for unordered image sets. In Proceedings of ECCV.
81. Schiller J H, Voisard A (2004) Location-based services. Morgan Kaufmann.
82. Schindler G, Brown M, Szeliski R (2007) City-scale location recognition. In Proceedings of IEEE CVPR.
83. Schindler G, Krishnamurthy P, Lublinerman R, Liu Y, Dellaert F (2008) Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In Proceedings of IEEE CVPR.
84. Serdyukov P, Murdock V, van Zwol R (2009) Placing Flickr photos on a map. In Proceedings of SIGIR.
85. Simon I, Seitz S M (2008) Scene segmentation using the wisdom of crowds. In Proceedings of ECCV.
86. Simon I, Snavely N, Seitz S M (2007) Scene summarization for online image collections, In Proceedings of IEEE ICCV.
87. Singh V, Gao M, Jain R (2010) Social Pixels: genesis and evaluation. In Proceedings of ACM Multimedia.
88. Snavely N, Garg R, Seitz S M, Szeliski R (2008) Finding paths through the world's photos. ACM Trans. Graph. 27(3).
89. Snavely N, Seitz S M, Szeliski R (2006) Photo Tourism: Exploring photo collections in 3D. ACM Trans. Graph. 25(3):835–846.
90. Snavely N, Seitz S M, Szeliski R (2008) Modeling the world from internet photo collections. Int. J. Comput. Vision 80(2):189–210.
91. Szeliski R (2005) Where am I? In Proceedings of IEEE ICCV Computer Vision Contest. http://research.microsoft.com/en-us/um/people/szeliski/VisionContest05/old_ideas.htm.
92. Sunkavalli K, Romeiro F, Matusik W, Zickler T, Pfister H (2008) What do color changes reveal about an outdoor scene. In Proceedings of CVPR.
93. Torniai C, Battle S, Cayzer S (2006). Sharing, discovering and browsing geotagged pictures on the web. Springer.
94. Toyama K, Logan R, Roseway A (2003) Geographic location tags on digital images. In Proceedings of ACM Multimedia.
95. Trinder J C, Wang Y (1998) Automatic road extraction from aerial images. Digital Signal Processing, 8 (4), 215-224.
96. Tsai C -M, Qamra A, Chang E (2005) Extent: Inferring image metadata from context and content. In Proceedings of IEEE ICME.
97. Tsikrika T, Diou C, de Vries A, Delopoulos A (2009) Image annotation using clickthrough data. In Proceedings of ACM CIVR.
98. Tuytelaars T, Van Gool L (2004) Matching widely separated views based on affine invariant regions. International Journal on Computer Vision.

99. Ueda T, Amagasa T, Yoshikawa M, Uemura S (2002) A system for retrieval and digest creation of video data based on geographic objects. In Proceedings of International Conference on Database and Expert Systems Applications.
100. Wei X -Y, Jiang Y –G, Ngo C –W (2009) Exploring inter-concept relationship with context space for semantic video indexing. In Proceedings of ACM CIVR.
101. Wolf L, Bileschi S (2006) A critical view of context. International Journal of Computer Vision 68(1):43–52.
102. Yanai K, Kawakubo H, Qiu B (2009) A visual analysis of the relationship between word concepts and geographical locations, In Proceedings of CIVR.
103. Yanai K, Yaegashi K, Qiu B (2009) Detecting cultural differences using consumer-generated geotagged photos. In Proceedings of International Workshop on Location and the Web.
104. Yu J, Luo J (2008) Leveraging probabilistic season and location context models for scene understanding. In Proceedings of ACM CIVR.
105. Yuan J, Luo J, Kautz H, Wu Y (2008) Mining GPS traces and visual words for event classification. In Proceedings of Multimedia Information Retrieval (MIR).
106. Zhang W, Kosecka J (2006) Image based localization in urban environments. In Proceedings of 3DPVT.
107. Zheng V W, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with GPS history data. In Proceedings of WWW.
108. Zheng Y, Wang L, Zhang R, Xie X, Ma W-Y (2009) GeoLife: managing and understanding your past life over maps. In Proceedings of MDM.
109. Zheng Y, Zhang L, Xie X, Ma W –Y (2009) Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of WWW.
110. Zheng Y, Zhao M, Song Y, Hartwig A, 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 CVPR.