Crowdsourcing What Is Where: Community-Contributed Photos as Volunteered Geographic Information

Shawn Newsam
University of California at Merced

Leveraging large collections of georeferenced, community-contributed photographs can help solve three knowledge-discovery problems: annotating novel images, annotating geographic locations, and performing geographic discovery.

The popularity of media-sharing services such as Flickr and YouTube has created large collections of community-contributed multimedia data whose growth shows no signs of slowing. These publicly available data sets represent a rich, but largely unstructured, source of information that's attracting attention from the multimedia analysis research community. The rich context of this social media (such as user-provided tags, comments, geolocations, time, and device metadata) not only supports traditional forms of knowledge discovery (such as learning the correspondences between the textual tags and the visual content), but also enables novel research directions that were inconceivable just a few years ago.

This article focuses on one such broad research direction, namely knowledge discovery based on the geographic location of social media. We focus in particular on large collections of georeferenced community-contributed photographs, such as those available at Flickr or Panoramio (see http://www.panoramio.com). These services let users specify the location of their shared images using the standard geographic convention of a longitude and latitude pair either manually through placement on a map or automatically using image metadata provided by a GPS-enabled camera.

Even though access to large collections of georeferenced images is a relatively new development, a growing body of work already exists within the multimedia research community investigating the use of this data for knowledge discovery. Because this often entails geographic knowledge discovery, a primary goal of this article is to connect this research thrust to the larger phenomenon of volunteered geographic information (VGI; see the ''Volunteered Geographic Information'' sidebar for more information). In particular, we argue that georeferenced social media **is** another form of VGI and, as such, the geographic discovery it enables is in effect crowdsourcing what is where on the earth's surface.

Considering community-contributed photo collections in a VGI context also helps us organize state-of-the-art research in this area and identify promising future directions. We survey existing work by grouping it into three broad classes of problems: leveraging georeferenced collections to annotate novel images, annotate geographic locations, and perform geographic discovery. We feel that while this last class of problems has received the least attention, it has great potential as an alternate to traditional means of geographic inquiry and truly recognizes georeferenced social media as a first-class VGI citizen.

Leveraging collections to annotate novel images

More work exists on leveraging georeferenced collections to annotate novel images than the other two classes of problems. The appeal of this problem is that it can help with the broader problem of managing large image collections, particularly personal collections. Methods have been developed for semantically annotating novel images with known locations,

Volunteered Geographic Information

In 2007, geographer Michael Goodchild coined the term *volunteered geographic information* (VGI)[1] to refer to the growing collections of geographically relevant information provided voluntarily by individuals. Enabled by emerging technologies centered around the Web, this phenomenon is creating sources of geographic information that differ along many dimensions from traditional sources. While some of these differences present challenges, such as the legitimacy of the contributors and the relative lack of provenance information, others are enabling large-scale geographic discovery not possible before in terms of reduced temporal latency and providing the ''people's'' perspective. Another important driving force behind VGI is that the data and derived products are usually made available through open source licenses. This can have simple (but profound) implications, particularly in regions where map data is covered by restrictive licenses or even censored.

The nature of VGI varies greatly. On one hand, the data and its presentation can be similar to traditional formats such as Google Maps, Yahoo! Maps, and Microsoft's Bing Maps. Perhaps the best example of this kind of VGI is OpenStreetMap , a ''free editable map of the whole world ... made by people like you'' (see http://www.openstreetmap. org). Awareness campaigns primarily in the form of mapping parties have boosted participation in OpenStreetMap so that its coverage already rivals that of commercial online maps.

At the other end of the VGI spectrum are projects such as the Audubon Society's Christmas Bird Count (see http:// www.audubon.org/Bird/cbc), which is now in its 110th year. During a three-week period each year, tens of thousands of volunteers throughout the US record and report on the number and types of birds they observe. Other distinctively nontraditional VGI projects include the pop-versus-soda website (http://www.popvssoda.com/) that maps local preferences for referring to carbonated beverages based on online surveys and the CommonCensus Map Project, which is ''redrawing the map of the United States ... to reveal the boundaries people themselves feel, as opposed to the state and county boundaries drawn by politicians'' (see http://www.commoncensus.org/). Maps are drawn based on spheres of influence, such as which major city people feel has the most cultural and economic influence on their area overall and which professional sports teams they support.

Clearly, VGI spans a range of information types and sources. Up to this point, however, no one has explicitly made the connection of how georeferenced social media can be considered another form of VGI.

Social media, particularly community-contributed multimedia data, represents a rich but complex source of volunteered information (and is thus receiving significant research attention, as evidenced by this special issue). Because much of this data is annotated with at least approximate location information, it can be interpreted in a geographic context. Goodchild makes this observation using georeferenced Flickr images of Ayer's Rock in Australia as an example.[1] However, his connection doesn't go beyond using the location information to organize individual images for map-based browsing.

On the other hand, computer science researchers working in multimedia analysis have recognized the potential for knowledge discovery in georeferenced online photo collections. Most of the work in this area, however, has not explicitly made the connection to VGI.

Reference

1. M.F. Goodchild, ''Citizens as Sensors: The World of Volunteered Geography,'' *GeoJournal,* vol. 69, no. 4, 2007, pp. 211-221.

as Figure 1a (next page) shows. This is particularly useful for images captured using GPS-enabled cameras, as the system-generated annotations allow the images to be organized and searched at a more meaningful level than with low-level image descriptors such as color and texture. Methods have also been developed for providing location annotations of novel images—that is, estimating where the images were acquired (see Figure 1b). Here, the locations are predicted using the images' visual content at times in combination with textual tags. While the annotation is performed in all cases by treating reference collections of georeferenced images as VGI, the goal of this first class of problems—annotating novel images—doesn't result in geographic discovery.

Semantic annotation

The goal of the works described in this section is to provide semantic-level annotations of novel images with known locations. The results indicate that, in general, location is a stronger cue than visual content, temporal information, and other metadata typically associated with the images.

SpiritTagger from Moxley et al.[1] is a tag suggestion tool for novel georeferenced images

(a)

Novel image with known location

Georeferenced images →
- Text tags
- Event/activity labels
- Region labels
- Identities of people
- Links to Wikipedia articles

(b)

Novel image whose location is unknown

Georeferenced images →
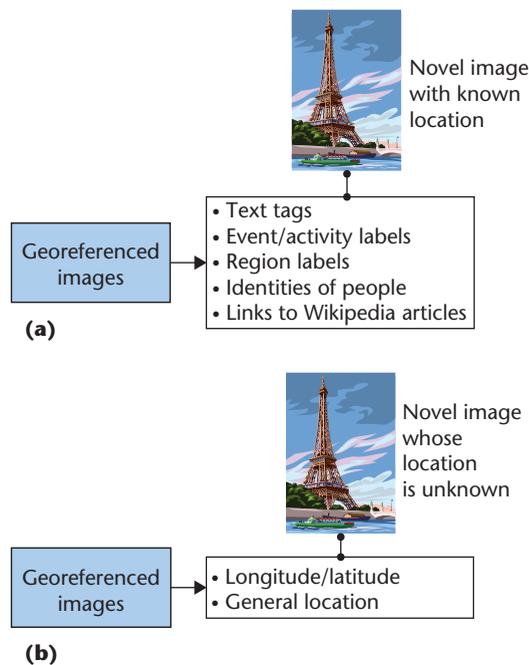- Longitude/latitude
- General location

Figure 1. (a) Reference sets of georeferenced images have been used to semantically annotate novel images with known locations. The annotations include text tags such as "Eiffel Tower," event/ activity labels such as "vacation," region labels such as "sky," identities of people appearing in the image, and links to relevant Wikipedia articles (for the Eiffel Tower, in this case). (b) Reference sets of georeferenced images have also been used to annotate the locations of novel images—that is, to estimate where in the world they were acquired. This can either be a precise longitude and latitude pair or a coarse assignment such as "Paris."

such as those acquired from a GPS-enabled camera. The tool uses the image's location to perform a geographically constrained visual-similarity search against a reference collection of georeferenced Flickr images. Textual tags from the top-ranked images become candidate annotations for the novel image. SpiritTagger bases the final ranking of the candidate tags on their local importance—tags that have a high local-to-global frequency of occurrence are ranked higher. This approach is effective for annotating a diverse set of images such as suggesting the tags "surfer," "wave," and "Santa Barbara" for a photograph of someone surfing in Santa Barbara, California, and the tags "baseball," "field," and "Angels" for a photograph of an Anaheim Angels baseball player during a game.

Several works go beyond tag propagation and seek to classify georeferenced images with a constrained set of labels. Joshi and Luo[2] label novel images with a set of 16 event/ activity descriptions such as "a visit to the beach" or "wedding." They apply separate classifiers to the image's location information and visual content. They use the image's location to generate a set of close-by geo-tags using a gazetteer. They then use a large reference set of Flickr images to learn the association between events/activities and gazetteer geotags in a probabilistic framework. They classify the image's visual content using established visual detectors from the video concept-detection community, and use late fusion to combine the outputs of the location and visual classifiers. The authors demonstrate that, in general, the gazetteer geotags, and thus the location information, is a stronger cue than the visual content for event and activity recognition—but the combination of both performs better than either separately.

Cao et al.[3] annotate groups of novel images at the event and scene level. A group of images is assigned one of 12 possible event labels such as "skiing" and "graduation," while the individual images are assigned one or more of 12 possible scene labels, such as "coast" and "kitchen." The authors use conditional random fields to model the hierarchical relationship between events and scenes as well as the relationship between the scene labels of nearby images with respect to time and location. They cluster novel images into groups using time and location information. Then they use belief propagation to simultaneously label a group of images collectively as an event and individually as scenes. Visual features feed into both the event and scene labeling. Application to a large collection of personal photographs shows that location information improves over time information alone, and that joint labeling at the event and scene levels improves over labeling the two levels independently.

Yu and Luo[4] label individual regions of a novel image with a constrained set of 11 tags, such as "grass," "sky," and "snow" using visual features in combination with coarse temporal information (the season in which the image was taken) and coarse location information (the US state in which the image was taken). They use a unified graphical model to learn the probabilistic dependencies between the various image characteristics—coarse acquisition time, coarse acquisition location, individual image region visual features, individual image

region labels (provided through manual labeling), and the spatial arrangement of the image regions—from reference collections of Flickr and consumer images. They demonstrate that even coarse location information improves region-labeling accuracy.

Naaman et al.[5] propose a system for annotating the identities of people appearing in personal photo collections where the time and location of the photos are known. The system interactively suggests candidate identity labels based on the photo's metadata, including its time and location, its membership in sets of photos related to events and locations, and the co-occurrence statistics of people already identified in the photo. While it performs no image analysis (such as face detection or recognition), the authors note that the system could easily be combined with automated image-understanding modules. Although they originally proposed the system for personal photo collections, it could also be applied to annotating the identities of celebrities or politicians on a larger scale.

Quack et al.[6] describe a system for linking novel images whose location is only approximately known—say at the city scale—to relevant Wikipedia articles. They cluster a reference set of georeferenced images based on their visual content and textual tags. Then they use key phrases identified through frequent item-set mining to find candidate Wikipedia articles for the clusters. They validate the articles by visually comparing the images in a cluster with images in the candidate articles. The system annotates a novel image by identifying one or more clusters containing visually similar images and then simply linking to the corresponding articles. They show that the system is effective for linking images of popular landmarks, such as the Arc de Triomphe or Notre Dame, to the appropriate Wikipedia articles.

Divvala et al.[7] investigate different context types for visual object detection, including geographic context. They leverage the IM2GPS system of Hays and Efros[8] to geolocate a novel image and assign it 15 geographic properties, such as land-cover probability, vegetation density, light pollution, and elevation-gradient magnitude based on maps of the estimated image location. The motivation is that object class occurrence is correlated with geography—for example, "boat" is frequently found in water regions and "person" is more likely to be located in densely populated regions. They demonstrate that while geographic context determined in this manner isn't as effective as other kinds of context for improving object detection, combining contexts performs better than any alone.

Location annotation

Collections of georeferenced photo collections have also been leveraged to annotate the location of novel images—that is, to estimate where in the world the photo was taken. Most systems typically achieve this through a visual similarity search against a set of georeferenced images at times in conjunction with textual tags.

While there's a history of work on the related problem of using image-to-image correspondences to register novel images to a constrained set of target images (including a location-recognition contest titled "Where Am I?" at the 2005 IEEE International Conference on Computer Vision that used images taken of a single city with a calibrated camera, and which included overlapping fields of view), perhaps the first work on the potentially more difficult problem of geolocating novel images on a larger spatial scale using community-contributed images is by Jacobs et al.[9] Their goal is to estimate the location of webcams distributed around the US. Rather than analyze the specifics of the scenes, they use principal component analysis to characterize image variations relating to the diurnal cycle and the weather. These variations are effective for either geolocating novel cameras through correlation to existing cameras with known locations or, more interestingly, through correlation to satellite images, including synthetic satellite images that simply map daylight. Results using a database of more than 17 million images from 538 static outdoor cameras located across the US show that the approach is able to localize most cameras to within 50 miles of the true location.

Hays and Efros[8] tackle the difficult problem of globally geolocating a single image using only its visual content. They use a reference set of 6.5 million georeferenced Flickr images derived from an initial set of 20 million images by removing images annotated with geographically uninformative tags, such as "birthday" or "camera phone." Their system geolocates novel images by performing nearest-neighbor

searches in the reference set using a comprehensive set of visual features. Results on a challenging set of test images of which around only 5 percent are of recognizable tourist sites show the approach is able to locate about a quarter of the images to within a small country (approximately 750 kilometers) of their true location, or about 30 times better than chance.

Gallagher et al.[10] extend the work by Hays and Efros[8] to incorporate textual tags in geolocating novel images. The reference data set here is a collection of 1.2 million georeferenced Flickr images compiled from the site's "interesting" images (see http://www.flickr.com/explore/interesting). Using the interest level as a filter helps ensure that the quality and content of the collection's images are reasonable. They compare three different approaches for geolocating a novel image with text annotations:

▌ a visual baseline that uses only visual content similar to Hays and Efros;[8]

▌ a tag baseline in which they use a tag-probability map derived from the reference set to geolocate the tags of the novel image; and

▌ a combined visual-plus-tag approach that filters reference images based on their tag intersection with the novel image and then ranks them based on visual similarity.

Results on a 2,000-image test set demonstrate that textual tags perform better than visual content—but they perform better in combination than alone.

Cao et al.[11] recognize the difficulty of estimating the exact location at which a photo was taken and instead estimate only the coarse location. They spatially cluster a reference set of georeferenced images with text annotations using the mean-shift algorithm. Then they use logistic canonical correlation regression to model the mapping between visual content and text annotations and the spatially disjoint clusters. The system georeferences a novel image by assigning it to the best cluster based on its visual content and annotations. This approach is effective when applied to a reference set of georeferenced Flickr images from across the US. In this case, mean-shift clustering results in 451 distinct regions across the US, where each region constitutes roughly 20,000 square kilometers on average, or a square region measuring 140 km on each side.

Crandall et al.[12] also only estimate the approximate location of a novel photo. They investigate geolocation at two scales, which they term metropolitan and landmark. Similar to Cao et al.,[11] they spatially cluster a reference set of georeferenced images with text annotations using the mean-shift algorithm. The system performs separate clusterings at the metropolitan and landmark scales by adjusting the mean-shift parameters. These clusters' statistics provide information on the most-photographed cities in the world and the most-photographed landmarks in a city and in the world. Then they use support vector machines to perform the mapping between visual content and text annotations and spatially disjoint clusters. They only do this for clusters corresponding to popular cities and landmarks. Their system also geolocates a novel image by assigning it to the best cluster based on its visual content and annotations.

Applying this approach using a reference data set of more than 30 million georeferenced Flickr images from across the world reaches different conclusions at different scales. At the landmark scale, both the text annotations and visual content perform better than chance, while at the metropolitan scale only the text annotations perform better than chance. This makes sense because of the relatively larger variation in visual content in images at the metropolitan scale. The system also uses the temporal information of a set of images by the same photographer to improve the individual images' geolocation accuracy. The insight here is that photographers typically take multiple pictures of the same landmark to capture different viewpoints, lighting conditions, subjects, and so on—thus, neighboring frames provide nonredundant visual evidence of where the photos were taken.

In Cristani et al.,[13] the system clusters a large set of georeferenced images using proximity and visual similarity concurrently. The resulting clusters, termed *geocategories*, are regions that are geographically coherent in that they contain images depicting a particular landscape such as mountains or coast. The system performs the clustering using probabilistic latent semantic analysis in which the geocategories are considered latent variables. Then the system geolocates a novel photo by computing its

most likely geocategory membership based on its visual characteristics. Interestingly, this method can geolocate sets of images that are colocated, and the authors demonstrate that this is significantly more effective than geolocating the images individually.

Leveraging collections to annotate geographic locations

Before, we described methods for leveraging georeferenced collections to annotate novel images primarily to aid in managing large image data sets. Now we can discuss methods for annotating geographic locations (as Figure 2a shows). This is a task in which social media is considered more explicitly as VGI, because the objective is more in line with the fundamental geographic problem of determining what is where on the earth's surface. Researchers have developed two types of annotation systems:

❚ *textual*, that identify representative tags (keywords or phrases) for the geographic regions; and

❚ *visual*, that identify representative images for the visually salient features of a region, such as landmarks.

Crandall et al.[12] visually annotate prominent landmarks. The spatial clustering of a collection of georeferenced images results in groupings of images that are near such landmarks. A random selection of images from these clusters is unlikely to result in a coherent set of representative images, so the system identifies one or more canonical images using visual analysis. The system poses canonical image selection as a graph-partitioning problem in which the images are the nodes and the edges indicate the similarity between pairs of images based on interest-point features. It uses spectral clustering to identify tightly connected clusters of photos. It chooses the images corresponding to the nodes with the largest weighted degrees in the resulting clusters as the canonical images for the landmarks. This approach to visually annotating geographic locations is effective for prominent landmarks across Europe and the US, leading the authors to conclude that ''while individual users of Flickr are simply using the site to store and share photos, their collective activity reveals a striking amount of geographic and visual information about the
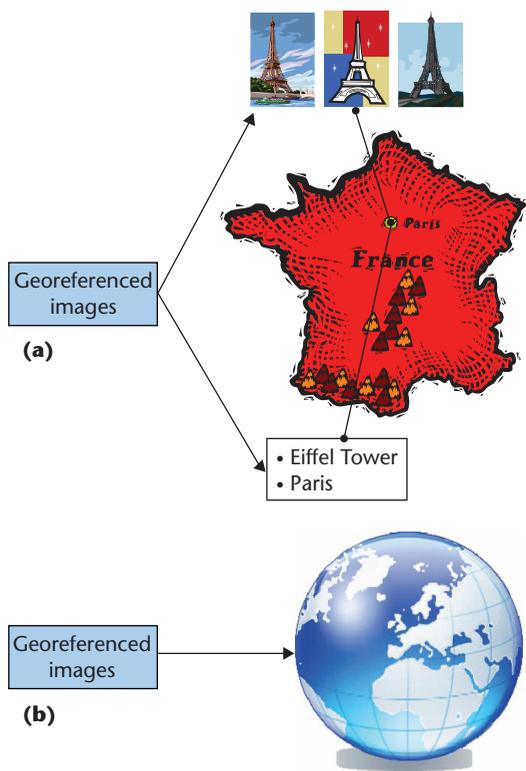

Figure 2. (a) Reference sets of georeferenced images have been used to annotate geographic locations. This includes textual annotation with representative tags and visual annotation with representative images. (b) Reference sets of georeferenced images have been used for geographic discovery. This includes detecting interesting cultural differences, discovering the most-photographed landmarks in the world, estimating weather satellite images, and producing land-cover maps.

world.''[12] This statement is the closest that any research has come to recognizing georeferenced social multimedia as a form of VGI.

Visual annotation of landmarks is performed on a larger, worldwide scale by Zheng et al.[14] They mined visually consistent images of landmarks using two methods. First, they clustered a large number of georeferenced images based on their locations. Then they clustered the resulting geoclusters with images contributed by more than a predetermined number of photographers based on their visual content. This results in approximately 14,000 visual clusters for 2,240 landmarks from 812 cities in 104 countries. The second method performs text mining on a large number of travel guide articles from an online travel site (see http://www.wikitravel.com) to identify candidate

landmark names. Then they perform a Google image search and visually cluster the resulting images. This results in approximately 12,000 visual clusters for 3,246 landmarks from 626 cities in 130 countries. There's surprisingly little overlap between the two result sets, and the combined list of landmarks consists of 5,132 unique landmarks from 1,259 cities in 144 countries.

Kennedy and Naaman[15] and Kennedy et al.[16] present a two-step process for identifying both representative tags and representative images for geographic locations. The first step, which they expand upon in follow-up work,[17] uses a large collection of georeferenced images with annotations to identify tags that correspond to locations, events, or spatially localized events. The insight is that location tags should exhibit distinctive spatial patterns, such as burstiness perhaps over multiple spatial scales, while event tags should exhibit distinctive temporal patterns. The system extracts these patterns from the location and time metadata from a large collection of annotated images.

The second step identifies representative images of the geographic locations signified by the representative tags using visual content. Given a set of images annotated with a location tag (typically a landmark), the system performs visual clustering to find common views of the landmark. Then it ranks these views (clusters of visually similar images) according to their representativeness. The system deems representative clusters to be those that contain images from many different users, are visually cohesive, and contain images that are distributed uniformly in time. Finally, the images corresponding to the highly ranked views are themselves ranked to identify the representative images for the geographic location. It deems representative images to be those that are visually similar to other images in the view, are visually dissimilar to random images from other views, and feature commonly photographed local structures for the view. They demonstrate this two-step representative tag- and image-discovery process on a large set of georeferenced Flickr images from the San Francisco area. They show that the visual annotations' accuracy for 10 popular landmarks benefits from both the visual analysis and the geographic constraints provided by the location tag-discovery step.

Chen et al.[18] take the process of using large collections of georeferenced images to annotate geographic locations one step further by automatically generating tourist maps showing popular landmarks as vectorized icons. Points of interest (POI) are identified by clustering the images using both location and textual annotations. Similar to others' work, their system identifies canonical images of the POIs—which are usually landmarks—using visual analysis. The system then computes homographies between the canonical images and uses them to derive a single consensus image for the POI. Finally, the system transforms the consensus image into a vectorized icon using tooning techniques. The final product is a tourist map containing the icons. They use this approach to generate maps of San Francisco and Rome from large collections of Flickr images.

Leveraging collections for geographic discovery

We now describe a third class of problems that treats social media as a form of VGI—namely leveraging georeferenced community-contributed photo collections for geographic discovery. The Merriam-Webster dictionary describes geography as "a science that deals with the description, distribution, and interaction of the diverse physical, biological, and cultural features of the earth's surface." Accordingly, we consider geographic discovery to be a process that derives knowledge about what is where on the earth's surface in the broad sense of the term "what." Simply put, we can use this information (as Figure 2b shows) to generate maps not only of the physical aspects of our world, such as the terrain, but also of the cultural and behavioral aspects. While there has been relatively little work in this area, we feel it has significant potential for realizing the full worth of georeferenced social media as VGI, particularly as an alternate to traditional means of geographic inquiry. We note that while the works attributed to this third class of problems have some overlap with the works in the other two classes, we feel they're set apart by their capacity for generating maps of phenomena often not observable through other means.

Yanai et al.[19] discover spatially varying cultural differences among concepts such as "wedding cake" from large collections of georeferenced images. The cultural differences are manifest in the concepts' appearance, which the proposed system detects using visual

analysis. First, it clusters a large number of images corresponding to a concept based on visual content. Then it discards the smaller clusters as well as those with low intracluster similarity. Finally, the system spatially clusters the remaining images and selects representative images of the concept for a region in an unsupervised fashion using a probabilistic latent semantic indexing framework. The authors then manually compare representative images for different regions to detect cultural differences among a concept's appearance. They apply the approach to a large collection of Flickr images and detect that wedding cakes tend to be simpler and smaller in Europe than in the US, as well as other interesting visual cultural variations. This work is an example of how we can use georeferenced social media to generate cultural maps. You can imagine how a similar visual- and spatial-clustering approach could be used to automatically map terms which have more divergent meanings, such as how the word "pop" is variously used to refer to either a carbonated beverage or male relative in different parts of the world.

While the primary objective of Crandall et al.[12] is to aid in organizing large collections of georeferenced images, as we previously described, the landmark- and metropolitan-scale spatial clustering of large collections of georeferenced images reveals interesting properties about popular cities and landmarks. The clustering empirically discovers what people consider to be the most significant landmarks both in the world and within specific cities, which cities are the most photographed, which cities have the highest and lowest proportions of attention-drawing landmarks, which views of these landmarks are the most characteristic, and how people move through the cities and regions as they visit different locations within them. Some of the results of applying the technique to a global data set of more than 30 million images are surprising. For example, the authors find the Apple Store in midtown Manhattan to be the fifth most-photographed place in New York City and the 28th most-photographed place in the world. This work demonstrates how we can use georeferenced social media to generate a behavioral map—in this case, of tourist hot spots.

Jacobs et al.[10] geolocate static cameras by finding the maximum correlation between ground-level and satellite images. They also consider the intriguing reverse question: can we use a collection of widely distributed webcams with known locations to derive a weather satellite image? They show this is feasible by using regularized linear regression to learn a mapping from a set of webcam images to satellite images in a supervised fashion. Then they use this mapping to generate a novel satellite image from a set of webcam images. They apply this method to a set of 42 cameras in the Maryland and Virginia area by using a training set of 1,400 weather satellite images. The predicted weather satellite images, which are essentially maps, are similar to the true images in terms of the cloud-cover patterns.

A particularly exciting new research direction that we term *proximate sensing* exploits georeferenced photos to automatically identify physical features on the earth's surface in much the same way that remotely sensed images, such as overhead images from satellite or aerial platforms, have been used for decades. One advantage of ground-level images is, however, their potential for discriminating between land-use classes. While remote sensing might be useful for deriving maps of land cover (which refers to the vegetation, structures, or other features that cover the land), it's much less effective at deriving maps of land use (which refers instead to how humans use land). Land parcels with different land uses— for example, a hospital and a shopping center—might share similar land cover (a building and a parking lot), and thus be difficult to distinguish in overhead imagery, especially in low-resolution imagery. Proximate sensing instead relies on ground-level images of close-by objects and events, and thus could resolve such ambiguities.

Research on proximate sensing, however, has so far focused only on land-cover mapping. The latent geocategories in the work by Cristani et al.[5] produce a map-like partitioning of a country-sized region into geographically coherent subregions. The geocategories turn out to correspond in large part to broad land-cover classes such as cities, fields, mountain areas, mountain villages, coastal areas, and lakes. A subsequent manual labeling of the geocategories could thus generate a land-cover map.

Our own work on proximate sensing[14] goes a step further and investigates whether we can use georeferenced images to perform land-cover classification in a completely automated

fashion. We classify individual ground-level images using support vector machines as depicting developed or undeveloped scenes based on their visual content. We then spatially aggregate these labels to generate two-class land-cover maps. We applied this approach to a 100×100-km region of the United Kingdom and evaluated it using ground-truth data.

We use two georeferenced photo collections to perform the land-cover classification. The first is a set of more than 100,000 community-contributed photos from the publicly available Geograph Britain and Ireland project. This project aims to ''collect geographically representative photographs and information for every square kilometer of Great Britain and Ireland'' (see http://www.geograph.org.uk). The second data set is nearly 1 million georeferenced Flickr images. While both of the predicted maps are similar to the ground truth, the results of using the Geograph images are more accurate, as would be expected because of the difference in the photographers' intent. We also show that classifiers learned in a weakly supervised fashion by propagating labels from the ground-truth maps to the training images actually outperformed classifiers learned using manually labeled images. This is significant, because generating manually labeled training sets is labor intensive.

Conclusion

We described three classes of multimedia analysis problems in which large collections of community-contributed images are treated as VGI. We used this grouping to present a snapshot of the state of the art of knowledge discovery in georeferenced social media. While we do not claim that this is an exhaustive survey, it provides a good overview of the work being performed in this nascent area.

The last class of problems—leveraging georeferenced image collections for geographic discovery—has significant potential as an alternate to traditional means of mapping the physical, cultural, and behavioral aspects of the earth's surface. This is especially true for phenomena that might otherwise be difficult to observe, such as the size of wedding cakes in different countries. There are plenty of opportunities to extend the initial work in this area—for example, to perform land-use classification—by incorporating recent advances in computer vision on object, scene, and event recognition.

We're interested in finding out just how effective georeferenced social multimedia is for crowdsourcing what is where on the earth's surface. **MM**
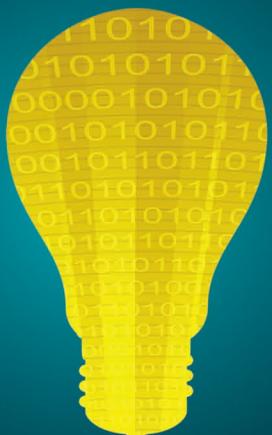
References

1. E. Moxley, J. Kleban, and B.S. Manjunath, ''Spirit-Tagger: A Geo-Aware Tag Suggestion Tool Mined from Flickr,'' *Proc. ACM Int'l Conf. Multimedia Information Retrieval,* ACM Press, 2008, pp. 24-30.
2. D. Joshi and J. Luo, ''Inferring Generic Activities and Events from Image Content and Bags of Geo-Tags,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* ACM Press, 2008, pp. 37-46.
3. L. Cao et al., ''Annotating Collections of Photos Using Hierarchical Event and Scene Models,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2008, pp. 1-8.
4. J. Yu and J. Luo, ''Leveraging Probabilistic Season and Location Context Models for Scene Understanding,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* IEEE CS Press, 2008, pp. 169-178.
5. M. Naaman et al., ''Leveraging Context to Resolve Identity in Photo Albums,'' *Proc. ACM/IEEE-CS Joint Conf. Digital Libraries,* ACM Press, 2005, pp. 178-187.
6. T. Quack, B. Leibe, and L. Van Gool, ''World-Scale Mining of Objects and Events from Community Photo Collections,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* ACM Press, 2008, pp. 47-56.
7. S. Divvala et al., ''An Empirical Study of Context in Object Detection,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2009, pp. 1271-1278.
8. J. Hays and A. Efros, ''IM2GPS: Estimating Geographic Information from a Single Image,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2008, pp. 1-8.
9. N. Jacobs et al., ''Geolocating Static Cameras,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2007, pp. 1-6.
10. A. Gallagher et al., ''Geo-Location Inference from Image Content and User Tags,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2009, pp. 55-62.
11. L. Cao et al., ''Enhancing Semantic and Geographic Annotation of Web Images via Logistic Canonical Correlation Regression,'' *Proc. ACM Int'l Conf. Multimedia,* ACM Press, 2009, pp. 125-134.
12. D. Crandall et al., ''Mapping the World's Photos,'' *Proc. Int'l World Wide Web Conf.,* ACM Press, 2009, pp. 761-770.

13. M. Cristani et al. ''Geo-Located Image Analysis Using Latent Representations,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2008, pp. 1-8.
14. Y.-T. Zheng et al., ''Tour the World: Building a Web-Scale Landmark Recognition Engine,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2009, pp. 1085-1092.
15. L. Kennedy and M. Naaman, ''Generating Diverse and Representative Image Search Results for Landmarks,'' *Proc. Int'l World Wide Web Conf.,* ACM Press, 2008, pp. 297-306.
16. L. Kennedy et al., ''How Flickr Helps Us Make Sense of the World: Context and Content in Community-Contributed Media Collections,'' *Proc. ACM Int'l Conf. Multimedia,* ACM Press, 2007, pp. 631-640.
17. T. Rattenbury and M. Naaman, ''Methods for Extracting Place Semantics from Flickr Tags,'' *ACM Trans. on the Web,* vol. 3, no. 1, 2009, pp. 1-30.
18. W.-C. Chen et al., ''Visual Summaries of Popular Landmarks from Community Photo Collections,'' *Proc. ACM Int'l Conf. Multimedia,* ACM Press, 2009, pp. 789-792.
19. K. Yanai, K. Yaegashi, and B. Qiu, ''Detecting Cultural Differences Using Consumer-Generated Geo-tagged Photos,'' *Proc. Int'l Workshop on Location and the Web,* ACM Press, 2009, pp. 1-4.
20. D. Leung and S. Newsam, ''Proximate Sensing: Inferring What-Is-Where from Georeferenced Photo Collections,'' *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* IEEE CS Press, 2010, pp. 1-8.

Shawn Newsam is an assistant professor of electrical engineering and computer science and a founding faculty member at the University of California at Merced. His research interests include knowledge discovery in complex data through computer vision and pattern recognition. Newsam has a PhD in electrical and computer engineering from the University of California at Santa Barbara. Contact him at snewsam@ucmerced.edu.

cn *Selected CS articles and columns are also available for free at http://ComputingNow. computer.org.*