Modeling People and Places with Internet Photo Collections

Understanding the world from the sea of online photos

David Crandall, School of Informatics and Computing, Indiana University

Noah Snavely, Department of Computer Science, Cornell University

Computational photography often considers sets of photos taken by a single user in a single setting, but the popularity of online social media sites has created a social aspect to photo collections as well. Photo-sharing sites such as Flickr and Facebook contain vast amounts of latent information about our world and human behavior. Our recent work has involved building automatic algorithms that analyze large collections of imagery in order to understand and model people and places at a global scale. Geotagged photographs can be used to identify the most photographed places on Earth, as well as to infer the names and visual representations of these places. At a local scale, we can build detailed three-dimensional models of a scene by combining information from thousands of two-dimensional photographs taken by different people and from different vantage points. One key representation for many of these tasks is a network: a graph linking photos by visual similarity or other measures.

This article describes our work in using online photo collections to reconstruct information about the world and its inhabitants at both global and local scales. This work has been driven by the dramatic growth of social content-sharing Web sites, which have created immense online collections of user-generated visual data. Flickr.com alone currently hosts more than 6 billion images taken by more than 40 million unique users,¹¹ while Facebook.com has said it grows by nearly 250 million photos every day.²⁰

While users of these sites are primarily motivated by a desire to share photos with family and friends, collectively they are generating vast repositories of online information about the world and its people. Each of these photos is a visual observation of what a small part of the world looked like at a particular point in time and space. It is also a record of *where* a particular person (the photographer) was at a moment in time and *what* he or she was paying attention to.

In aggregate, and in combination with nonvisual metadata available on photo-sharing sites (including photo timestamps, geotags, captions, user profiles, and social contacts), these billions of photos present a rich source of information about the state of the world and the behavior of its people. We can thus imagine extending the domain of computational photography to encompass all of the world's photos, where the goal is to extract useful information about places and people from our collective image data.

We have recently demonstrated how to reconstruct information about the world at both global and local scales using collections such as those on Flickr.^{7,9,21} Our algorithms for analyzing the world at a global scale can automatically create annotated world maps by finding the most photographed cities and landmarks, inferring place names from text tags, and analyzing the images themselves to identify “canonical” images to summarize each place. Figures 1 and 2 show examples of such maps.

Figure 1 is an annotated map of North America, automatically generated by analyzing nearly 35 million photos from Flickr. For each of the top 30 most photographed cities, the map shows

An Annotated Map of North America

Figure 9, shows a similar “point cloud” reconstruction of the Old Town of Dubrovnik, created from more than 6,500 Flickr photos, as well as a reconstruction of the Forum from a larger 3D model of Rome.

Automatically-Generated Annotated Map of Europe

The challenge is that online data sets are largely unstructured and thus require sophisticated algorithms that can organize and extract meaning from noisy data. In our case, this involves developing automated techniques that can find patterns across millions of images. Representing large image collections as networks or graphs, where each image is a node and is connected to related images, can form a useful representation for extracting many types of information, such as 3D structures or representative views. This observation suggests interesting parallels between image collections and other domains where link structure appears, such as between people in social networks or between pages in the World Wide Web.

FIGURE 3

A 3D Model of a Scene Is Generated from Photographs

Figure 6 shows an example of a visual connectivity network for a set of images of Trafalgar Square. We compute a measure of visual similarity between every pair of images and connect those above a threshold. Many photos are not connected at all; they generally are images of people or objects and not of the square itself. A clustering algorithm finds tightly connected components of the network. This produces three groups of images, each corresponding to a different frequently photographed scene from the square (marked by a dotted blue line in the figure).

MAPPING THE WORLD

In addition to the images themselves, modern photo-sharing sites such as Flickr collect a rich assortment of nonvisual information about photos. Many online photos have metadata specifying what a photo contains (text tags), as well as where (geotag), when (timestamp), and how (camera metadata such as exposure settings) the photo was taken. On social media sites photos are also

FIGURE 4

Metadata for a Photo Downloaded from Flickr

accompanied by information generated as a result of sharing, such as text tags, comments, and ratings. The geotagging features of photo-sharing sites are especially useful in our work. These geotags record the latitude and longitude of the point on Earth where a photo was taken. This information either is entered manually by the photographer using a map-based interface or (increasingly) is automatically determined by a GPS (global positioning system) receiver in the camera or cell phone. Figure 4 shows example metadata for a photo downloaded from Flickr,¹⁵ including a geotag specifying latitude and longitude, textual tags, and camera information.

By aggregating this visual and nonvisual information from the photographs of many millions of users, we can study what the world looks like in the collective consciousness of the world's photographers. To begin, we collected a data set of more than 90 million geotagged photos using the Flickr public API.² As one might expect, more photographs are taken in some locations than others. A plot of the geotags in our database, shown in figures 1 and 2, illustrates this nonuniform distribution. This distribution contains significant information beyond the images themselves that is revealed only through analysis of large numbers of photos from many photographers. For example, photo-taking is dense in urban areas and quite sparse in most rural areas. Note that the continental boundaries in these maps are quite sharp, because beaches are such popular locales to take photos. Also note how roads are visible in these maps because people take photos as they travel. In figure 1 the east-west interstate highways crossing the western United States are especially clear.

Given that photographic activity is highly nonuniform, we identify geographic concentrations of photos by using mean shift, a clustering algorithm for finding the peaks of a non-parametric distribution.⁵ We look for peaks at multiple scales (by applying mean shift with kernels of different sizes), including both city (50-km radius) and landmark (100-m) scales. We can then rank cities and landmarks based on the number of photos or number of distinct photographers who have uploaded a photo from that place.

For example, the most photographed cities in the world, according to Flickr, are New York, London, San Francisco, Paris, and Los Angeles. The five most photographed landmarks are the

FIGURE 5

SIFT Feature Matching

(a) input photograph

(b) input photo with extracted SIFT features superimposed

(c) a pair of overlapping images shown side-by-side with matching features connected by black lines

Eiffel Tower, Trafalgar Square, Tate Modern, Big Ben, and Notre Dame (more detailed rankings are available online⁸). The techniques used to produce these rankings are relatively simple, but they are an example of the kinds of analyses that are suddenly possible with the rise of photo-sharing sites. The list of top landmarks includes some surprises; the Apple Store in Manhattan, for example, ranks among the top five landmarks in New York City and is ranked 28th in the entire world.

For each of these highly photographed places, we can automatically infer its name by looking at the text tags that people assign to photographs taken in that place. Although most tags are at best weakly related to geography—flower, family, sunset, blackandwhite, etc.—we can find place names by looking across the photos of millions of users and finding tags that are used frequently in a particular place and infrequently outside of it. We can also generate a visual description of each place by finding a representative image that summarizes that place well. To do this, we deem each photograph taken in a place as a vote for the most interesting scene at that location. Intuitively, we then try to find the scene that receives the most votes by looking for groups of photos that are visually similar and taken by many different users.

To implement this intuitive approach we construct a graph in which each image from the place is a node, and we connect pairs of photos having a high degree of visual similarity. Then we apply a graph-clustering algorithm to find tightly connected components of the graph (i.e., groups of nodes that are connected to many other nodes within the group but not to many nodes outside the group)

FIGURE 6

Photo Network for Finding Representative Images

and choose one of these photos as a representative image. A sample graph of this type is shown in figure 6. To decide which nodes to connect, we measure visual similarity using an automated technique called SIFT (scale-invariant feature transform) feature matching,¹⁴ illustrated in figure 5. Note that this summary image is not necessarily the *best* photo of a particular place—it will likely be a canonical tourist photo rather than a more unusual yet captivating viewpoint captured by a professional photographer.

The map in figure 1 was produced completely automatically using this analysis on tens of millions of images downloaded from Flickr. Starting with a blank slate, we plotted the raw photo geotags to produce the map in the background and then applied mean-shift clustering to locate the 30 most photographed cities on earth. For each of those cities, we extracted the city's name by looking for distinctive text tags and found the name of the most photographed landmark within the city. Then we extracted a representative image for that landmark. While the analysis is not perfect—a human would have chosen a more appropriate image of Phoenix than a bird on a baseball field, for example—the result is a compelling summary of North America, produced automatically by analyzing the activity of millions of Flickr users. Maps for other continents, regions, and cities of the world are available at our project Web site.⁸

This analysis is reminiscent of sociologist Stanley Milgram's work during the 1970s studying people's "psychological maps"—their mental images of how the physical world is laid out.¹⁷ He asked Parisians to draw freehand maps of their city and then compared these maps with the factual geography. Milgram found that the maps were highly variable and largely inaccurate but that most

people tended to anchor their maps around a few key landmarks such as the River Seine and Notre Dame Cathedral. He ranked landmarks by their degree of importance in the collective Parisian psychology by counting the number of times that each landmark was mentioned in the study. Our work is an analogous study, at a much larger scale. It is important to note that we are also dealing with much less controlled data, however, and our results are biased by the demographics of Flickr users.

Data from Flickr can also be used to study the behavior of human photographers, because each photo is an observation of what a particular user was doing at a particular moment in time. For example, studying sequences of geotagged, timestamped photos can track the paths that people take. Figure 7 shows an example of this analysis for Manhattan. Note that the grid structure of the streets and avenues is clearly visible, as are popular tourist paths such as the walk across the Brooklyn Bridge and the ferries leaving the southern tip of the island. We used this data to study the relationship between human mobility patterns and the social network defined on Flickr.⁶ We can infer a user's social network with startling accuracy based only on such patterns. After observing that two people were at about the same place at about the same time on five distinct occasions, for example, the probability that they are friends is nearly 60 percent.

This illustrates how data from online social-sharing systems can be used to study questions in sociology at a scale that has never before been possible. It also reveals a potential privacy concern with geotagging-enabled social-networking sites, in that users can reveal more information about themselves than they intend to, such as the identity of their friends.

FIGURE 7

Trails of Human Movement in Manhattan

RECONSTRUCTING LANDMARKS

Thus far, our visual representation of a landmark has simply been a single image that is visually similar to many other images taken at that site. For popular landmarks, however, thousands of online photos are taken by different users, each with a different composition and from a different viewpoint. Each of these photos is thus a slightly different 2D observation of a 3D scene. This leads to the idea of using computer vision algorithms to recover 3D geometry from these photos completely automatically.

We developed a technique to reconstruct accurate 3D models of world landmarks from large collections of uncalibrated images on photo-sharing Web sites.^{1,9} The principle underlying this technique is similar to that used by stereopsis, which allows humans to perceive the world in 3D. Our two eyes view a scene from slightly different perspectives, and from these two views the brain can infer the depth of each scene point based on the difference between where the point appears in the two images. The corresponding computer vision problem of inferring depth given the input from two different cameras is the well-studied stereo problem.¹²

In the case of reconstructing landmarks using Flickr images, there are not two, but thousands, of images serving as independent views. The problem is much more difficult, however, because the precise positions and viewing directions of the cameras are not known. (The latitude-longitude coordinates in geotags are much too noisy for this purpose. Even the geotags produced by GPS receivers are very noisy because consumer GPS devices have an accuracy of about 10 meters.) Hence, both the structure of the scene and the positions of all of the cameras must be inferred simultaneously. This is known in computer vision as the *structure from motion* problem. While structure from motion has been studied for decades, Internet photo collections pose new challenges to computer vision because of their extreme scale and unstructured nature—they are taken by many different people with many different cameras, from largely unknown viewpoints. Moreover, the images on a site such as Flickr contain significant noise arising from mislabeled images, poor-quality photos, image occlusions, and transitory objects (such as people) appearing in the scene.

Solving this problem means first knowing which images of a given landmark have visual overlap. As with our technique for choosing representative views, we first perform SIFT feature matching between pairs of images to build an image network. Unrelated images, such as a closeup of a pigeon, are automatically discarded, as they will not be connected to other images that actually feature the landmark. This matching algorithm is computationally expensive but is easily parallelized.⁹

Figure 5 illustrates SIFT feature matching in more detail. Given the input photo (a) on the top left of figure 5, SIFT extracts a number of features, consisting of salient locations and scales in the image, as well as a high-dimensional descriptor summarizing the appearance of each feature. A subset of detected feature locations, depicted as yellow circles, is superimposed on the image (b) on the top right. The image is shown again on the bottom (c) next to an image from a similar viewpoint; we can match SIFT features to find a correspondence between these images. Because of the robustness of SIFT, most of these matches are correct.

Once we have the network of visual connectivity between images, we need to estimate the precise position and orientation of the camera used to capture each image—that is, exactly where the photographer was standing and in which direction he or she was pointing the camera—as well as the 3D coordinates of every SIFT point matched in the images. It turns out that this can be posed as an enormous optimization problem, in which the location of each scene point and the position of

each camera are estimated given constraints induced by the same scene points appearing in multiple images. This optimization tries to find the camera and scene geometry that, when related to each other through perspective projection, most closely agree with the 2D SIFT matches found between the images. This optimization problem is difficult to solve, not only because of its size, but also because the objective function is highly nonlinear.

Information in the visual network, as well as absolute location information from geotags, can help with this reconstruction task, however. Consider a pair of visually overlapping images, such as the two photographs shown in the upper left of figure 8. Using the computed SIFT matches and geometric reasoning algorithms, we can determine the geometric relationship between these two images—that image 2 is taken to the left of image 1 and rotated slightly clockwise, say. We can compute such *relative camera poses* for each edge in the network (such as the small network on the right of figure 8). By computing many such relationships, we can build up a network of information on top of a set of images, shown to the right. We also have geotags for some images, shown as latitude/longitude coordinates. Unfortunately, these geotags are very noisy and can at times be

hundreds of meters away from a photo's true location. On the other hand, some geotags are quite accurate. If we knew which were good, we could propagate locations from those photos to their neighbors in the network. Given a raw set of geotagged photos, however, we do not know which are accurate. To overcome this problem we have developed a new technique that uses the image network to combine these position estimates in a more intelligent, robust way, “averaging out” errors in the noisy observations by passing geometric information between nodes in the image network. This algorithm uses a *message-passing* strategy based on a technique known as *loopy belief propagation* commonly used in machine learning, computer vision, and other areas.¹⁸ This algorithm is scalable and can find good solutions to very nonlinear problems. While complex, our algorithm starts with the simple idea that each image should repeatedly average its location with that of its neighbors, hence using the graph to smooth noisy location estimates. Because of the extreme noise, this simple averaging approach doesn't work well; therefore, we developed a more sophisticated approach.⁹

The message-passing process described here repeats for a number of rounds, so each image repeatedly updates its position based on information from its neighbors. This algorithm results in fairly accurate camera positions, and applying standard optimization techniques (such as gradient descent) using these positions as a starting point can yield further improvements. With this algorithm we have built some very large 3D models, including the reconstructions of the city of Dubrovnik and parts of Rome shown in figure 9. To process these large problems, we implemented the algorithm using the MapReduce framework and ran these as jobs on a large Hadoop cluster. (For more information, see our project's Web page¹⁰). In other work on the 3D modeling problem, we reconstructed all of the major sites in Rome from hundreds of thousands of Flickr photos in less than 24 hours (thus reconstructing “Rome in a Day”).^{3,19}

While photo-sharing sites such as Flickr and Facebook continue to grow at a breathtaking pace, they still do not have enough images to reach our eventual goal of reconstructing the entire world in 3D. The main problem is that the geospatial distribution of photographs is highly nonuniform, as noted in the last section—there are hundreds of thousands of photos of Notre Dame but virtually none of the café across the street.

FIGURE 9

3D Reconstructions

old town of Dubrovnik

the Forum, Rome

One solution to this problem is to entice people to take photos of underrepresented places through *gamification*. This is the idea behind PhotoCity, an online game developed in collaboration with the University of Washington. In PhotoCity, teams of players compete against one another by taking photos at specific points in space to capture flags and buildings.²² Through this game, we collected more than 100,000 photos of the Cornell and University of Washington campuses over a period of a few weeks. We used these photos to reconstruct large portions of the two campuses, including areas that otherwise did not have much photographic coverage on sites such as Flickr. A few example building models created from these photos, along with a screenshot of the PhotoCity interface, are shown in figure 10. On the left is a screenshot of the PhotoCity interface showing an overhead map depicting the state of the game. On the right are a few 3D models created from photos uploaded by players.

Creating a successful game involved two key challenges: (a) building a robust online system for users to upload photos for processing; and (b) designing the game mechanics in such a way that users were excited about playing. To address the first challenge, we built a version of our 3D reconstruction algorithm that could take a new photo of a building and quickly integrate it into our current 3D model of that building, updating that model with any new information contributed by that photo.

For the second challenge of designing effective game mechanics, we developed a mix of incentives. One set of incentives involved competition at different levels (e.g., between students at the same school, as well as a race for each school to build the best model). Another set involved giving each player visual feedback about how much he or she contributed to the model, by showing 3D points created by that player's photos and by updating models so that players could see the progress of the game as a whole over time. A survey of players after the conclusion of the competition revealed that different players were motivated by different incentives; some were driven by competition, while others simply enjoyed seeing the virtual world grow over time.

FUTURE WORK

This article has presented some of our initial work on unlocking the information latent in large photo-sharing Web sites using network-analysis algorithms, but the true promise of this type of analysis is yet to be realized. The opportunities for future work in this area lie along two different lines. First, new algorithms are needed to extract visual content more efficiently and accurately: the algorithms presented here produce incorrect results on some specific types of scenes, for example, and they are relatively compute-intensive, requiring many hours on large clusters of computers to process just a few thousand images.

Second, this type of analysis could be applied to other disciplines. Many scientists are interested in studying the world and how it has changed over time, including archaeologists, architects, art historians, ecologists, urban planners, etc. As a specific example, the 3D reconstruction technique could simplify mapping remote archaeological sites,⁴ where using traditional laser range scanners is expensive and challenging. A cheaper and simpler alternative would be to use a digital camera to take many photos of a site, and then run our reconstruction algorithms on those photos once the researchers return from the field. As another example, we have recently studied how to automatically mine online photo collections for images of natural phenomena like snowfall and flowering, potentially giving ecologists a new technique for collecting observational data at a

FIGURE 10

PhotoCity

continental scale.²³

Imagine all of the world's photos as coming from a "distributed camera," continually capturing images all around the world. Can this camera be calibrated to estimate the place and time each of these photos was taken? If so, we could start building a new kind of image search and analysis

tool—one that would, for example, allow a scientist to find all images of Central Park over time in order to study changes in flowering times from year to year, or that would allow an engineer to find all available photos of a particular bridge online to determine why it collapsed. Gaining true understanding of the world from the sea of photos online could have a truly transformative impact.

An earlier version of this paper was presented as a keynote talk at Arts | Humanities | Complex Networks—a Leonardo satellite symposium at NetSci2010 (<http://artshumanities.netsci2010.net>).

REFERENCES

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R. 2009. Building Rome in a day. International Conference on Computer Vision.
2. The App Garden. 2010; <http://www.flickr.com/services/api/>.
3. Building Rome in a day; <http://grail.cs.washington.edu/projects/rome/>.
4. Chen, X., Morvan, Y., He, Y., Dorsey, J., Rushmeier, H. 2010. An integrated image and sketching environment for archaeological sites. Workshop on Applications of Computer Vision in Archaeology.
5. Comaniciu, D., Meer, P. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(5).
6. Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Science* 107(52):22436-22441.
7. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J. 2009. Mapping the world's photos. International World Wide Web Conference.
8. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J. Mapping the world's photos; <http://www.cs.cornell.edu/~crandall/photomap/>.
9. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D. 2011. Discrete-continuous optimization for large-scale structure from motion. Conference on Computer Vision and Pattern Recognition: 3001-3008.
10. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D. 2011. Discrete-continuous optimization for large-scale structure from motion; <http://vision.soic.indiana.edu/disco/>.
11. Flickr. 2011. 6,000,000,000; <http://blog.flickr.net/en/2011/08/04/6000000000/>.
12. Hartley, R., Zisserman, A. 2003. *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press.
13. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M. 2009. Computational social science. *Science* 323(5915):721-723.
14. Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal in Computer Vision*.
15. Melki, S. Photo reproduced under a Creative-Commons-Attribution license. Flickr user sergemelki; <http://www.flickr.com/photos/sergemelki/3391168464/in/photostream/>.
16. Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., Lieberman-Aiden, E. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176-182.
17. Milgram, S. 1976. Psychological Maps of Paris. In *Environmental Psychology: People and Their*

- Physical Settings*, ed. H. M. Proshansky, W. H. Ittelson, and L. G. Rivlin, 104-124. New York: Holt, Rinehart and Winston.
18. Pearl, J. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, California: Morgan Kaufmann Publishers.
 19. Photo Tourism; <http://phototour.cs.washington.edu/>.
 20. Shaffer, J. 2011. Bigger, faster photos. The Facebook Blog; <http://blog.facebook.com/blog.php?post=10150262684247131>.
 21. Snavely, N., Seitz, S., Szeliski, R. 2008. Modeling the world from Internet photo collections. *International Journal of Computer Vision* 80(2).
 22. Tuite, K., Snavely, N., Hsiao, D.-Y., Tabing, N., Popovi, Z. 2011. PhotoCity: training experts at large-scale image acquisition through a competitive game. Conference on Human Factors in Computing (CHI).
 23. Zhang, H., Korayem, M., Crandall, D., LeBuhn, G. 2012. Mining Photo-sharing Websites to Study Ecological Phenomena. International World Wide Web Conference.

LOVE IT, HATE IT? LET US KNOW

feedback@queue.acm.org

DAVID CRANDALL is an assistant professor in the School of Informatics and Computing at Indiana University in Bloomington, IN. He received a Ph.D. in computer science from Cornell University in 2008 and M.S. and B.S. degrees in computer science and engineering from Pennsylvania State University, University Park, in 2001. He was a postdoctoral research associate at Cornell from 2008-2010, and a senior research scientist with Eastman Kodak Company from 2001-2003. His research interests are computer vision and data mining, with a focus on visual object recognition, image understanding, machine learning, and mining and modeling of complex networks.

NOAH SNAVELY is an assistant professor of computer science at Cornell University, where he has been on the faculty since 2009. He received a B.S. in computer science and mathematics from the University of Arizona in 2003 and a Ph.D. in computer science and engineering from the University of Washington in 2008. Snavely works in computer graphics and computer vision, with a particular interest in using vast amounts of imagery from the Internet to reconstruct and visualize the world in 3D, and in creating new tools for enabling people to capture and share their environments. His thesis work was the basis for Microsoft's Photosynth, a tool for building 3D visualizations from photo collections. He is the recipient of a Microsoft New Faculty Fellowship and an NSF CAREER Award and has been recognized by *Technology Review's* TR35.

© 2012 ACM 1542-7730/12/0400 \$10.00