

NETWORKS OF PHOTOS, LANDMARKS, AND PEOPLE

David Crandall, School of Informatics and Computing, Indiana University, Bloomington, IN 47403, E-mail: <djcran@indiana.edu> Noah Snavely, Department of Computer Science, Cornell University, Ithaca, NY 14853, E-mail: <snavely@cs.cornell.edu>

Submitted: August 9, 2010

Abstract

Social photo-sharing sites like Flickr contain vast amounts of latent information about the world and human behavior. We describe our recent work in building automatic algorithms that analyze large collections of imagery in order to extract some of this information. At a global scale, we show how geo-tagged 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 show that we can build detailed 3-d models of a scene by combining information from thousands of 2-d photographs taken by different people and from different vantage points.

The dramatic growth of social content sharing websites has created immense collections of user-generated visual data online. Flickr.com alone currently hosts over 4 billion images taken by more than 40 million unique users [1], while Facebook.com grows by nearly 3 billion photos every month [2]. 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 their 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 the non-visual metadata available on photo sharing sites (including photo timestamps, geo-tags, 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.

In recent work, we have shown how vast photo collections like Flickr can be used to reconstruct information about the world at both global and local scales [3,4]. At a global level, we can create annotated maps of the world completely automatically, using the geo-tags on photos to reconstruct land boundaries, using tags to infer place names, and using visual analysis to find frequently photographed scenes (an example is shown in Figure 1). We can also use this analysis to generate statistics about places, such as ranking landmarks by their popu-

larity or studying which kinds of users visit which sites. At a more local level, we can use techniques from computer vision to automatically produce strikingly accurate 3-D models of a landmark, given a large number of 2-D photos taken by many different users from many different vantage points (see Figure 2).

This work is part of a larger emerging research trend within computer science that is studying how to use publicly available data from online social networking sites to address questions in a range of fields in the humanities and social sciences [5]. Compared to traditional techniques like surveys and direct measurement, data collection from online social networking sources is of negligible cost and can be conducted at unprecedented scales. The challenge is that online data is largely unstructured, thus requiring 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.

In this paper, we describe our recent work in using online photo collections to reconstruct the world at both global and local scales.

Fig. 1. 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 the name of the city inferred from tags, the name of the most photographed landmark, and a representative photo of the landmark. (© David Crandall)





Fig. 2. From an unstructured collection of photographs downloaded from Flickr (left), we can produce a 3-d model of the original scene (right). (© Noah Snavely)

Mapping the world

In addition to the images themselves, modern photo-sharing sites like Flickr also collect a rich assortment of non-visual information about photos. This information includes metadata recorded by the digital camera – exposure settings and timestamps, for example – as well as information generated as a result of the social sharing process – text tags, comments, and ratings, for example. Recently, photo sharing sites have introduced geo-tag features which record the latitude-longitude coordinates of where on Earth a photo was taken. This information is either entered manually by the photographer using a map-based interface, or (increasingly) comes directly from a Global Positioning System (GPS) receiver in the camera or cell phone. Many online photos thus include a rich variety of non-visual metadata, giving information about what a photo contains (text tags), as well as where (geotag), when (timestamp), and how (camera metadata) the photo was taken.

By aggregating this visual and non-visual 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 do this, we have collected a dataset of more than 90 million geotagged photos from Flickr using its public API [6]. Simply by plotting the geotags of these photos on a blank frame buffer, we can see how the distribution of photographs over the Earth’s surface is highly non-uniform, as shown in Figures 1 and 3. 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. The east-west interstate highways crossing the western United States in Figure 1 are especially clear.

Given that photographic activity is highly non-uniform, we identify geographic concentrations of photos by using Mean Shift [7], a clustering algorithm for finding the peaks of a non-parametric distribution. 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 number of photos or number of distinct photographers who have uploaded a photo from that place.

We find, for example, that the top most photographed cities in the world according to Flickr are New York, London, San Francisco, Paris, and Los Angeles, while the five most photographed landmarks on Earth are the Eiffel Tower, Trafalgar Square, Tate Modern, Big Ben, and Notre Dame. (See [8] for more detailed rankings.) The techniques we use to produce these rankings are relatively simple, but they are an example of the kinds of analysis that are suddenly pos-

sible 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 five top landmarks in New York City, and is ranked #28 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. While most tags are unrelated 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 also generate a visual description of each place by finding a “representative image” that summarizes that place well. To do this, we view 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. Comparing the visual content of two images is difficult and an active area in computer vision research; see [8] for the details of our approach.

Figure 1 shows a map produced completely automatically using the above 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

Fig. 3. Distribution of geo-tagged Flickr photos in Europe. (© David Crandall)



those cities, we extracted the city's name by looking for distinctive text tags, and also 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 striking 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 website [9].

This analysis is reminiscent of sociologist Stanley Milgram's work during the 1970s in studying people's "psychological maps" – their mental images of how the world is arranged [10]. He asked Parisians to draw freehand maps of their city, and then he compared these maps to the factual geography. He found that the maps across different people were highly variable and largely inaccurate, but that most people tended to anchor their maps around a few key landmarks like the river Seine and Notre Dame Cathedral. He then 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 can be thought of as analogous study, but at a much larger scale (and with less experimental control – our results are undoubtedly 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, by studying sequences of geo-tagged, time-stamped photos, we can track the paths that people take as they travel around a space. Figure 4 shows an example of this analysis for Manhattan. Note that the grid structure of the streets and avenues is clearly visible, as is popular tourist paths like the walk across the Brooklyn Bridge and the ferries leaving the southern tip of the island.

Reconstructing landmarks

In the work described so far, our visual representation of a landmark was simply a single image that was visually similar to many other images taken at that site. However, for popular landmarks there are thousands of online photos taken by different users, each with a different composition and from a different viewpoint. Each of these photos is thus a

slightly different two-dimensional observation of a three-dimensional scene.

We have developed a technique that can use photos from sharing websites to reconstruct accurate 3-d models of many landmarks [11]. The principle underlying the technique is similar to that used by stereopsis – the process that allows humans to perceive the world in 3-d. Our two eyes view a scene from slightly different perspectives, allowing our brains to infer the depth of a 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 has been studied extensively [12]. In the case of reconstructing landmarks using Flickr images, we have not two but thousands of images that serve as 2-d observations of a scene. However the problem is much more difficult because the precise positions and viewing directions of the cameras are not known ahead of time [13]; thus both the structure of the scene and the positions of all of the cameras must be inferred simultaneously. Moreover, the images on a site like Flickr contain significant noise, caused by factors like mislabeled images, poor-quality photos, image occlusions, and transitory objects (like people) appearing in the scene.

Reconstructing a landmark thus involves solving 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 is performed using a technique called incremental bundle adjustment [14], which works by creating an initial reconstruction using a small set of images and then iteratively adding more to the solution, rejecting images that are inconsistent (in order to be robust to noise and outliers).

Figure 2 shows an example reconstruction of the Colosseum in Rome, while Figure 5 presents reconstructions for several other major landmarks. More examples are available online at the Photo Tourism project website [15]. This technique was also recently used to reconstruct most of the popular tourist attractions of the entirety of Rome, in a completely automatic process that took under 24 hours [16].

While photo-sharing sites like Flickr and Facebook continue to grow at a breathtaking pace, there are still not enough images on these sites to reach our eventual goal of reconstructing the



Fig. 4. Trails of human movement in Manhattan, inferred from time-stamped, geo-tagged Flickr photos. Reprinted from [3], Figure (C) © 2009 International World Wide Web Conference Committee.

entire world in 3-d. The main problem is that the geospatial distribution of photographs is highly non-uniform, as we saw in the last section – there are hundreds of thousands of photos of Notre Dame, but virtually none of the café across the street. One possible solution to this problem is to explicitly entice users to take photos of under-represented places. This is the idea behind PhotoCity, an online capture-the-flag-like game in which teams of players compete against one another by taking photos at specific points in space [17]. The photos produced by this game have been used to reconstruct portions of the campuses of the University of Washington and Cornell University – areas which otherwise did not have much photographic coverage.

Future work

We have presented some initial work into unlocking the information latent in large photo-sharing websites, but the true promise of this type of analysis is yet to be realized. There are opportunities for future work in this area along two different lines. First, we need to develop new algorithms that can extract visual content more efficiently and accurately: the algorithms we present 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, we would like to find applications of this type of analysis to work in 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 3-d reconstruction technique could simplify mapping remote archaeological sites [18], 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.

References and Notes

This paper was presented as a keynote talk at Arts I Humanities I Complex Networks – a Leonardo satellite symposium at NetSci2010. See

<http://artshumanities.netsci2010.net/>

1. 4,000,000,000 (2009), <http://blog.flickr.net/en/2009/10/12/4000000000/>, accessed 1 August 2010
2. Faster, Simpler Photo Updates (2010), <http://blog.facebook.com/blog.php?post=206178097130>, accessed 1 August 2010.
3. D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg, "Mapping the World's Photos," in *International World Wide Web Conference*, 2009.
4. N. Snavely, S. Seitz, R. Szeliski, "Modeling the World from Internet Photo Collections," *International Journal of Computer Vision* **80**, No. 2 (November 2008).
5. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, "Computational Social Science," *Science* **323** No. 5915 (6 February 2009).
6. The App Garden (2010), <http://www.flickr.com/services/api/>, accessed 1 August 2010.
7. D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** No. 5 (May 2002).
8. Crandall [3].
9. Mapping the World's Photos, <http://www.cs.cornell.edu/~crandall/photomap/>

10. Milgram, S. "Psychological Maps of Paris," in *Environmental Psychology: People and Their Physical Settings*, (New York: Holt, Rinehart and Winston, 1976), pp. 104–124.
11. Manning [1] pp. 3-7.
12. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, (Cambridge University Press, 2003).
13. The latitude-longitude coordinates in geo-tags are much too noisy for this purpose. Even the geo-tags produced by GPS receivers are very noisy because consumer GPS devices have an accuracy of about 10 meters.
14. Snavely [4].
15. Photo Tourism, <http://phototour.cs.washington.edu/>.
16. S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski, "Building Rome in a Day," in *International Conference on Computer Vision*, 2009.
17. K. Tuite, N. Snavely, D.-Y. Hsiao, A. Smith and Z. Popović, "Reconstructing the World in 3D: Bringing Games with a Purpose Outdoors," in *International Conference on the Foundations of Digital Games*, 2010.
18. X. Chen, Y. Morvan, Y. He, J. Dorsey, H. Rushmeier, "An Integrated Image and Sketching Environment for Archaeological Sites," in *Workshop on Applications of Computer Vision in Archaeology*, 2010.

Fig. 5. Reconstruction results for a variety of different landmarks. Top row, left to right: St. Basil's Cathedral, Chartres Cathedral, Colosseum, Hagia Sofia. Second row: Half Dome, Notre Dame Cathedral, Pantheon, Michelangelo's Pietà. Third row: Mount Rushmore, Great Sphinx of Giza, Statue of Liberty, Stonehenge. Bottom row: St. Peter's Basilica, Trafalgar Square, Fontana di Trevi, Venus de Milo. (© Noah Snavely)

