3D exploitation of large urban photo archives

Citation: Peter Cho, Noah Snavely, and Ross Anderson (2010). 3D exploitation of large urban photo archives. Proc. SPIE 7697: 769714/1-8. ©2010 COPYRIGHT SPIE--The International Society for Optical Engineering.

As Published: http://dx.doi.org/10.1117/12.849767

Publisher: SPIE

Persistent URL: http://hdl.handle.net/1721.1/58581

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

3D Exploitation of Large Urban Photo Archives*

Peter Cho[a], Noah Snavely[b] and Ross Anderson[a]
[a]MIT Lincoln Laboratory, 244 Wood St., Lexington, MA, USA 02420
[b]Department of Computer Science, 4130 Upson Hall, Cornell University, Ithaca, NY 14853

ABSTRACT

Recent work in computer vision has demonstrated the potential to automatically recover camera and scene geometry from large collections of uncooperatively-collected photos. At the same time, aerial ladar and Geographic Information System (GIS) data are becoming more readily accessible. In this paper, we present a system for fusing these data sources in order to transfer 3D and GIS information into outdoor urban imagery. Applying this system to 1000+ pictures shot of the lower Manhattan skyline and the Statue of Liberty, we present two proof-of-concept examples of geometry-based photo enhancement which are difficult to perform via conventional image processing: feature annotation and image-based querying. In these examples, high-level knowledge projects from 3D world-space into georegistered 2D image planes and/or propagates between different photos. Such automatic capabilities lay the groundwork for future real-time labeling of imagery shot in complex city environments by mobile smart phones.

Keywords: Computer vision, 3D reconstruction, ladar georegistration, knowledge transfer.

1. INTRODUCTION

The quantity and quality of urban digital imagery are rapidly increasing over time. Millions of photos shot by inexpensive electronic cameras in cities can now be accessed via photo sharing sites such as Flickr and Google Images. However, imagery on these websites is generally unstructured and unorganized. It is consequently difficult to relate Internet photos to one another as well as to other information within city scenes. Some organizing principle is needed to enable intuitive navigating and efficient mining of vast urban imagery archives.

Fortunately, 3D geometry provides such an organizing principle for images taken at different times, places and resolutions. Recent work in computer vision has demonstrated automatic capability to recover relative geometric information for uncooperatively collected images [1]. Moreover, several other urban data sources including satellite views, aerial ladar point clouds and GIS layers can serve as absolute geometrical underlays for such reconstructed photo collections. High-level knowledge such as building names and city object geocoordinates can then transfer from the underlays into the ground-level imagery via geometrical projection.

In this paper, we demonstrate the utility of 3D geometry for enhancing 2D urban imagery following the flow diagram in Figure 1. Working with uncooperatively collected photos shot in New York City (NYC), we first reconstruct their cameras' relative positions and orientations as well as sparse urban scene geometry. We next fuse aerial ladar, satellite imagery and GIS data to produce a dense three-dimensional NYC map. The reconstructed photos are then georegistered with the 3D map. We illustrate good qualitative alignment between the independent ground-level and aerial data sets and estimate a quantitative registration error.

Once a 3D framework for analyzing 2D photos is established, many challenging urban image enhancement problems become tractable. We focus in particular on automatic feature annotation and image-based querying. Examples of geometry-mediated labeling of buildings and measuring of ranges to target points are presented. Finally, we close by summarizing our results and discussing future applications of this work.

* This work was sponsored by the Department of the Air Force under Air Force Contract No. FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Signal Processing, Sensor Fusion, and Target Recognition XIX, edited by Ivan Kadar,
Proc. of SPIE Vol. 7697, 769714 · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.849767

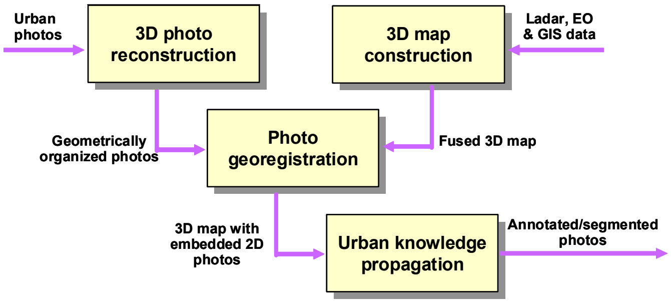
Urban photos
3D photo reconstruction
3D map construction
Ladar, EO & GIS data
Geometrically organized photos
Photo georegistration
Fused 3D map
3D map with embedded 2D photos
Urban knowledge propagation
Annotated/segmented photos

Fig 1: Algorithm flow diagram.

2. 3D PHOTO RECONSTURCTION

We begin by downloading more than 1000 digital photos shot of the lower Manhattan skyline and the Statue of Liberty from www.flickr.com. Flickr's website contains vast numbers of such pictures which users have tagged as generally related to New York City. But our uncooperatively collected data set is otherwise unorganized. We therefore next recover 3D structure from these photo collections using the SfM approach of Snavely *et al.* [1].

In brief, the 3D reconstruction procedure first extracts salient local features from each input image using Scale Invariant Feature Transform (SIFT) features, which are designed to be invariant to common geometric and photometric variations [2]. SIFT features can be robustly matched between all pairs of images via nearest-neighbor searching [2-4] plus RANSAC filtering [5].

SIFT feature matching itself begins to impose structure upon an unorganized set of photos. In particular, feature matches define an image connectivity graph [6]. Each photo corresponds to a node in the graph, and edges link images with matching SIFT features. The image graph for 1012 Manhattan photos is shown in Figure 2, using a graph layout and interactive rendering tool developed by M. Yee [7]. For this particular set of 1000+ photos, the graph divides roughly into two large clusters, representing the Manhattan skyline and the Statue of Liberty.

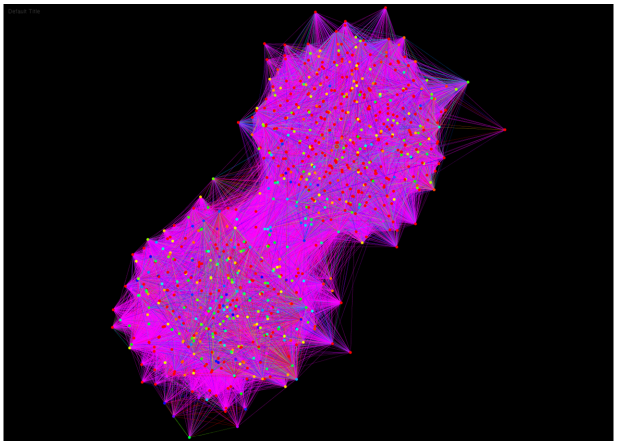

Fig 2: Graph illustrating topology for 1012 NYC photos. Nodes correspond to photos and are colored according to the estimated uncertainty in the corresponding camera location. SIFT feature overlap between photos is indicated by the graph's colored edges.

Yee's viewer enables one to intuitively inspect the graph's substructure. For example, if we zoom into the lower-left cluster of fig. 2, the circular nodes are replaced by thumbnails of Statue of Liberty photos (see fig. 3a). Similarly, skyline photos are grouped together within the upper-right cluster. Nodes located in the graph's neck region between the two large clusters exhibit both Statue and skyline content as one would expect (see fig. 3b).

Once corresponding features have been extracted and matched between multiple photos, our system next employs structure from motion techniques to recover camera poses and sparse scene structure. SfM takes as input the 2D feature matches and computes a set of 3D scene points, as well as the rotation, position and focal length parameters for each photo. We use the Bundler toolkit to solve the underlying optimization problem [8]. SfM results for the 1012-image collection of NYC Flickr photos are illustrated in Figure 4.

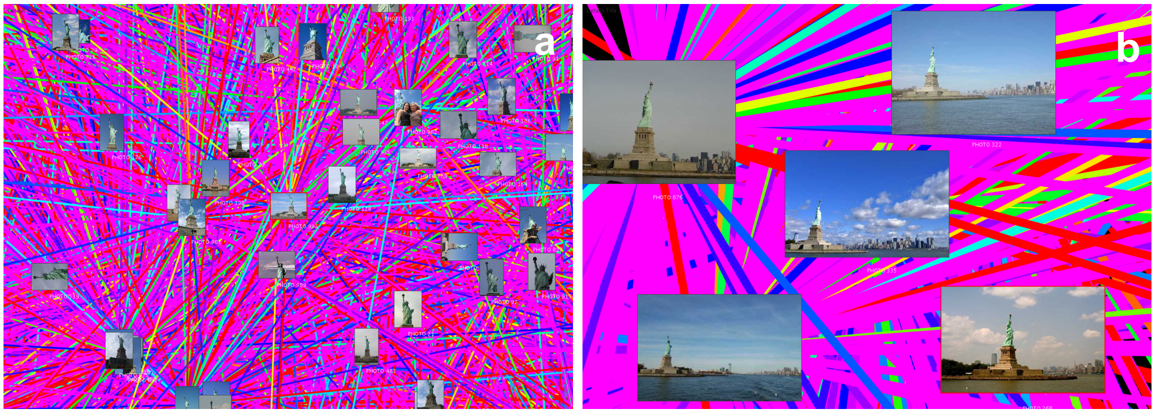

Fig 3 (a) Circular nodes in the lower-left subcluster of Figure 2 turn into Statue of Liberty photo thumbnails when the user zooms into the graph. (b) Thumbnails in the graph's middle neck region exhibit both Statue and skyline content.

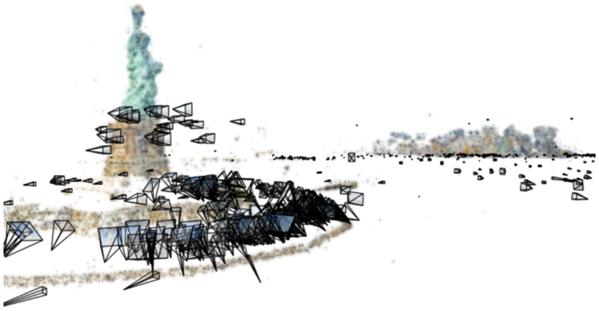

Fig 4. Relative positions and poses for 1012 cameras automatically determined by structure from motion. Relative statue and skyline geometry are also returned within a sparse point cloud output.

Given its high computational complexity, 3D reconstruction for large numbers of photos must currently be performed on a parallel cluster. We ran parallelized versions of the feature extraction and matching steps on Lincoln Laboratory's high-performance, 130-processor Grid [9]. The reconstruction process took approximately 4 hours on this cluster.

Conventional digital cameras only capture angle-angle projections of the 3D world onto 2D image planes. In the absence of metadata, photos yield neither absolute lengths nor absolute distances. It is therefore difficult to automatically determine absolute position, orientation or scale from a set of Internet images. In order to georegister reconstructed photos, we need to incorporate additional sensor data. We therefore next construct 3D urban maps based upon aerial ladar data.

3. 3D MAP CONSTRUCTION

High-resolution ladar imagery of entire cities is now routinely gathered by platforms operated by government laboratories as well as commercial companies. Airborne laser radars collect hundreds of millions of city points whose geolocations are efficiently stored in and retrieved from multiresolution quadtrees. Ladars consequently yield detailed geospatial underlays onto which other sensor measurements can be draped.

We work with a Rapid Terrain Visualization map collected over New York City on Oct 15, 2001. These data have a 1 meter ground sampling distance. By comparing absolute geolocations for landmarks in this 3D map with their counterparts in other geospatial databases, we estimate these ladar data have a maximum local georegistration error of 2 meters.

Complex urban environments are only partially characterized by their geometry. They also exhibit a rich pattern of intensities, reflectivities and colors. So the next step in generating an urban map is to fuse an overhead image with the ladar point cloud. For our New York example, we obtained Quickbird satellite imagery which covers the same area as the 3D data. Its 0.8 meter ground sampling distance is also comparable to that of the ladar imagery.

We next introduce GIS layers into the urban map. Such layers are commonplace in standard mapping programs which run on the web or as standalone applications. They include points (e.g. landmarks), curves (e.g. transportation routes) and regions (e.g. political zones). GIS databases generally store longitude and latitude coordinates for these geometrical structures, but most do not contain altitude information. Fortunately, height values can be extracted from the ladar underlay once lateral GIS geocoordinates are known.

After fusing together the ladar map, satellite image and GIS data, we derive the 3D map of New York City presented in Figure 5. In this map, the hue of each point is proportional to its estimated altitude, while saturation and intensity color coordinates are derived from the satellite imagery. GIS annotations supply useful context.

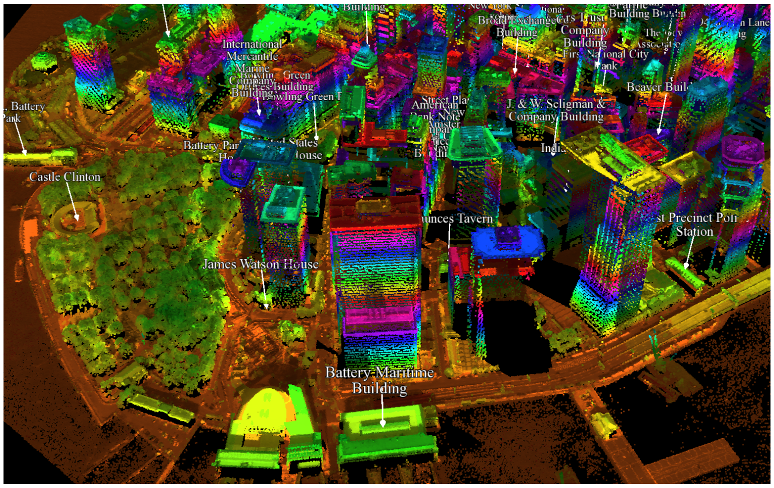

Fig 5. Fused 3D map of New York City.

Three-dimensional urban maps serve as global backdrops into which information localized in space and/or time may be incorporated. We therefore proceed to combine relative photo reconstructions with absolute coordinates from the NYC map to georegister large numbers of photos.

4. RECONSTRUCTED PHOTO GEOREGISTRATION

In order to georegister the SfM reconstruction (in its own relative coordinate system) with the absolute 3D urban map, we select 10 photos with large angular coverage and small reconstruction uncertainties. We then manually pick 33 features in the ladar map coinciding primarily with building corners and identify 2D counterparts to these features within the 10 photos. A least-squares fitting procedure subsequently determines the global transformation parameters needed to align all 1012 reconstructed photos with the ladar map.

This manual step could potentially be automated given GPS information for a subset of photos [10]. As future work, it would also be interesting to rerun the SfM optimization with constraints derived from correspondences between the SfM model and the ladar point cloud. This would help correct for any non-rigid deformations between the two data sources.

Figure 6 illustrates the reconstructed photos georegistered with the NYC map. In order to efficiently display large numbers of pictures in our 3D viewer, they are rendered as low-resolution thumbnails inside view frusta when the virtual camera is located far away in space. When the user double clicks on some view frustum, the virtual camera zooms in to look at the full-resolution version of the selected image. For example, Figure 7 illustrates a Statue of Liberty photo in front of the reconstructed Statue point cloud (for which we do not have ladar data). By comparing geocoordinates for reconstructed points on the Statue of Liberty with their pixel counterparts in Google Earth satellite imagery, we estimate that the average angular orientation error for our georegistered cameras is approximately 0.1 degree.

A more stringent test of the georegistration accuracy is provided by the alignment between projected ladar points and their corresponding image pixels, particularly for cameras located far away from their subject (e.g., the images of the skyline within Figure 8). Figure 9 exhibits the match between one skyline photo and the ladar background. Their agreement represents a nontrivial georegistration between two completely independent data sets. Similar qualitatively good alignment holds for nearly all of the skyline photos and the 3D map.

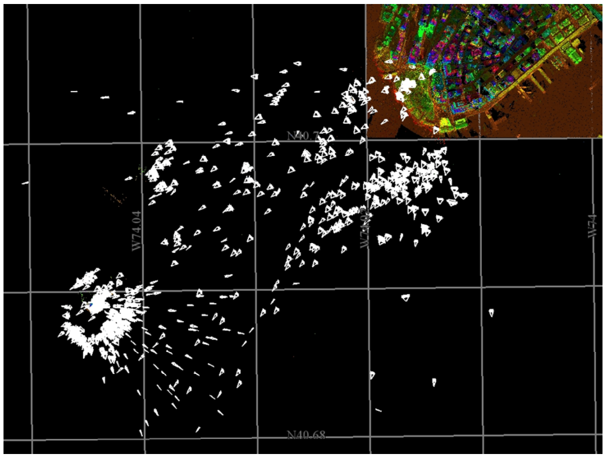

Fig 6. 1012 reconstructed photos georegistered with the 3D NYC map.

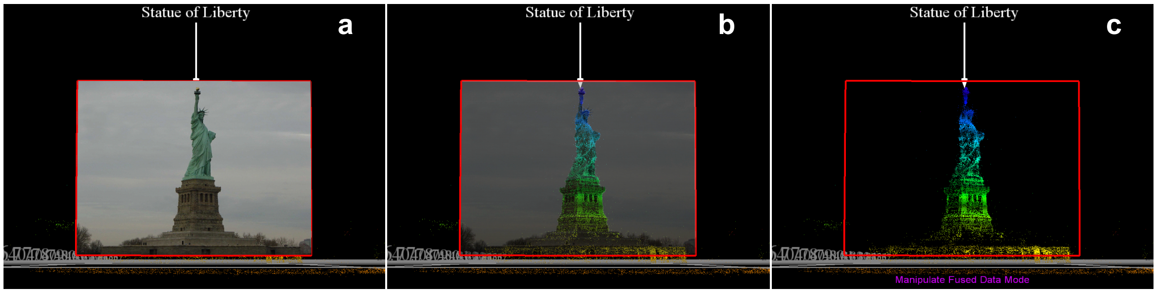

Fig 7. One Statue photo displayed in front of its reconstructed point cloud with (a) 0%, (b) 50% and (c) 100% alpha blending.

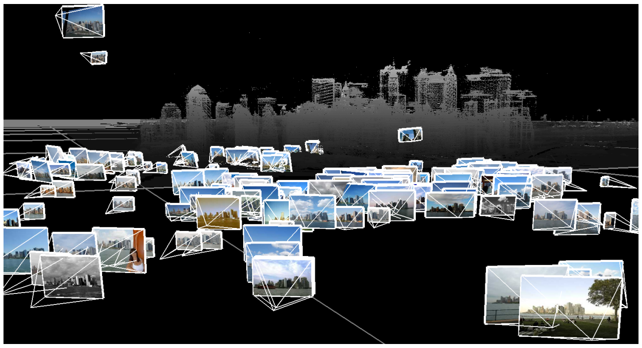

Fig 8. Approximately 500 reconstructed skyline photos georegistered with the 3D NYC map.

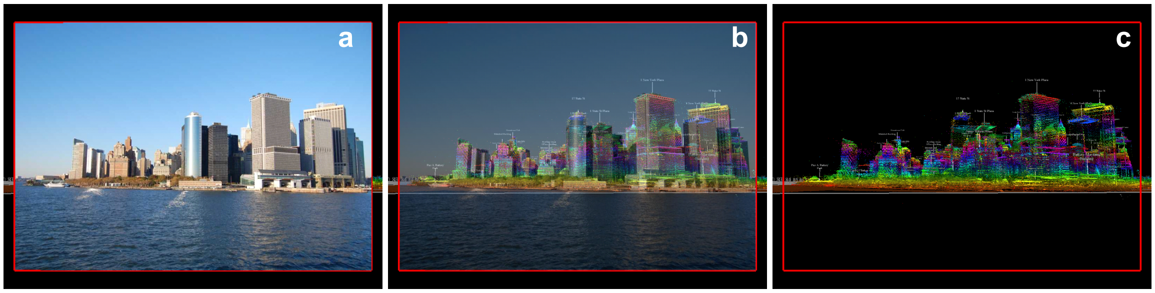

Fig 9. Alignment between skyline photo and 3D NYC map with (a) 0%, (b) 50% and (c) 100% alpha blending.

5. URBAN KNOWLEDGE PROPAGATION

Once the reconstructed photo collection is georegistered with the 3D urban map, many difficult image segmentation and enhancement problems become fairly straightforward. In this section, we present two proof-of-concept examples of geometry-based augmentation of geospatially organized photos which would be very difficult to perform via conventional computer vision algorithms.

Urban feature annotation

Our first enhancement application is automatically annotating static features in complex urban scenes. For example, we would like a machine to label buildings in NYC skyline photos. This annotation problem is extremely challenging from a conventional 2D standpoint due to the wide range of possible viewing and illumination conditions. But once a photo collection is georegistered, we leverage the fact that building names are tied to specific geolocations. After a camera has been globally reconstructed, projecting skyscraper geolocations into its image plane is simple. Skyscraper labels thus transfer automatically from 3D down to 2D. This basic projection approach holds for other information which is geospatially anchored, including roadway networks and political zones. Kopf *et al* present a similar annotation application in [11] which works on single photos. As in [1], our system can label large collections of photos all at once.

One technical problem for urban knowledge projection arises from line-of-sight occlusion. To overcome this issue, we convert our ladar point cloud into a height map and assume walls drop straight downwards from rooftop ladar data. If a ray traced from a 3D point back to a reconstructed camera encounters a wall, the point is deemed to be occluded from the camera's view, and information associated with that point is not used to annotate the image. We note that this approach works only for static occluders like buildings and not for transient occluders such as people and cars. Representative building name annotation results from this projection and raytracing procedure are shown in Figure 10.

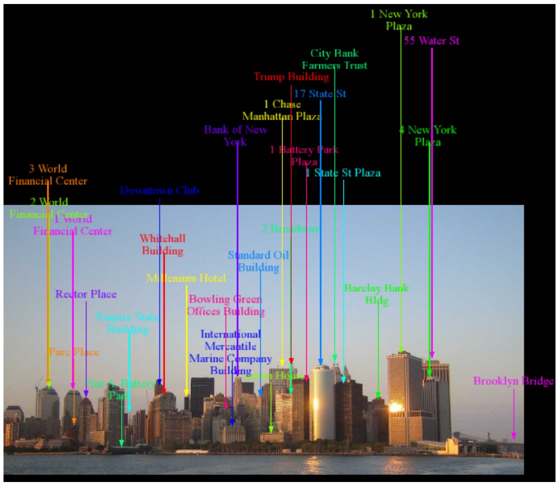

Fig 10. Skyline photo automatically annotated by projecting building names from the 3D NYC map.

Information transfer between images

Our second example demonstrates knowledge propagation between image planes mediated by intermediary geometry. Figure 11 illustrates a prototype image-based querying tool which exhibits the georegistered NYC photos in one window and a particular camera's view in a second window. When a user selects a pixel in the photo on the left, a corresponding 3D point is identified via raytracing in the map on the right. A set of 3D crosshairs marks the world-space counterpart. The geocoordinates and range for the raytraced world-space point are returned and displayed alongside the picked pixel in the 2D window. Note that ocean pixels selected in Figure 11 are reported to lie at 0 meter altitude above sea level, as expected

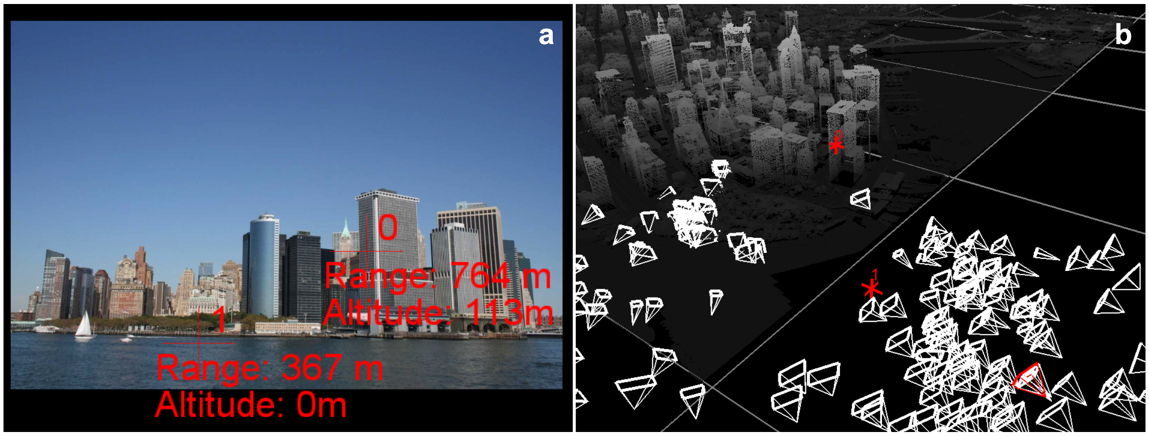

Fig 11. Image-based querying. (a) User selects 2 features in a georegistered photo. Machine subsequently raytraces these features back into the 3D map. (b) Points corresponding to selected photo pixels. Point ranges and altitudes above sea-level are displayed inside the photo window.

Once a 3D point corresponding to a selected 2D pixel is identified, it can be projected into any other camera so long as raytracing tests for occlusion are performed. For instance, the distances from a new camera to previously selected urban features of interest are reported in Figure 12. Similarly, static counterparts in overlapping air and ground views could be automatically matched. Even handoff of dynamic urban movers between multiple cameras should be possible provided the movers' length scales are *a priori* known.

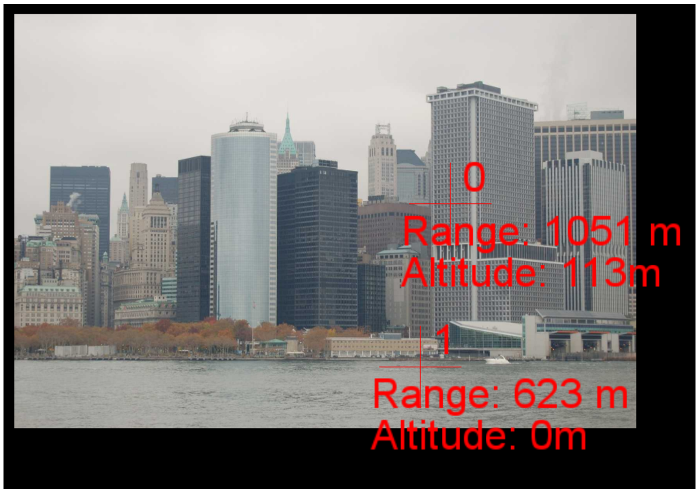

Fig 12. Points in fig 11b reprojected onto pixels in another photo. Camera ranges to features depend upon image, while urban feature altitudes remain invariant

6. SUMMARY AND FUTURE WORK

In this paper, we have demonstrated a prototype capability to reconstruct, georegister and enhance large numbers of uncooperatively collected urban digital photos. 3D photo reconstruction yields structured output from unstructured input. Ladar map registration augments photos with precise absolute geocoordinates and orientation metadata. And geometrical organization enables intuitive navigating and searching of large imagery archives.

Looking into the future, we foresee real-time interpretation of pixel outputs from mobile cameras operating in urban environments. As one walks, drives or flies through busy cities, information associated with buildings and streets could be used to automatically enhance an instantaneous image. Indeed, it should be possible to perform image-based queries of urban knowledge databases using photograph regions rather than text strings as inputs if accurate camera calibration can be rapidly calculated [12].

We are currently working with over 30,000 ground photos shot around the MIT campus in order to develop algorithms for mobile exploitation of urban imagery. Given its order-of-magnitude increase in size as well as its containing significant urban canyon occlusions, this next data set poses major new challenges beyond those discussed here for our initial 1000+ NYC photo experiments. But if technically feasible, near real-time annotation of live photo inputs would endow smart phones with powerful augmented urban reality capabilities. Given the dramatic rise in the quantity of digital images along with camera phones over the past few years, the impact of systematically leveraging vast numbers of photos could someday rival that of Google text search.

REFERENCES.

[1] Snavely, N., Seitz, S., and Szeliski, R., "Photo tourism: exploring photo collections in 3D," in *SIGGRAPH Conf. Proc.*, pp. 835-846 (2006).

[2] Lowe, D., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, (2004).

[3] Arya, S., Mount, D., Netanyahu, N., Silverman, R., and Wu, A., "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM),* vol. 45, pp. 891-923, (1998).

[4] Arya, S. and Mount, D., "Approximate Nearest Neighbor Searching," in *SODA*, pp. 271-280 (1993).

[5] Fischler, M. and Bolles, R., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* vol. 24, pp. 381-395, (1981).

[6] Snavely, N., Seitz, S., and Szeliski, R., "Skeletal graphs for efficient structure from motion," in *CVPR*, pp. 1-8 (2008).

[7] Sherrill, D., Yee, M., and Cho, P., "Analyzing networks of static and dynamic geospatial entities for urban situational awareness," in *SPIE*, p. 734605 (2009).

[8] Bundler, http://phototour.cs.washington.edu/bundler/

[9] Bliss, N., Bond, R., Kepner, J., Kim, H., and Reuther, A., "Interactive Grid Computing at Lincoln Laboratory," *Lincoln Laboratory Journal,* vol. 16, pp. 165-216, (2006).

[10] Kaminsky, R., Snavely, N., Seitz, S., and Szeliski, R., "Alignment of 3D Point Clouds to Overhead Images," in *CVPR*, pp. 63-70 (2009).

[11] Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., and Lischinski, D., "Deep photo: Model-based photograph enhancement and viewing," *SIGGRAPH Asia Conf. Proc.,* vol. 27, (2008).

[12] Takacs, G., Chandrasekhar, V., Gelfand, N., Xiong, Y., Chen, W., Bismpigiannis, T., Grzeszczuk, R., Pulli, K., and Girod, B., "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *MIR*, pp. 427-434 (2008).