

## MIT Open Access Articles

*Image-based querying of urban knowledge databases*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Cho, Peter, Soonmin Bae, and Fredo Durand. "Image-based querying of urban knowledge databases." Signal Processing, Sensor Fusion, and Target Recognition XVIII. Ed. Ivan Kadar. Orlando, FL, USA: SPIE, 2009. 733614-12. © 2009 SPIE--The International Society for Optical Engineering

**As Published:** <http://dx.doi.org/10.1117/12.818164>

**Publisher:** The International Society for Optical Engineering

**Persistent URL:** <http://hdl.handle.net/1721.1/52662>

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

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



# Image-Based Querying of Urban Knowledge Databases\*

Peter Cho<sup>a</sup>, Soonmin Bae<sup>b</sup> and Fredo Durand<sup>b</sup>

<sup>a</sup>MIT Lincoln Laboratory, 244 Wood St., Lexington, MA, USA 02420

<sup>b</sup>MIT Computer Science & Artificial Intelligence Laboratory, 32 Vassar St., Cambridge, MA 02139

## ABSTRACT

We extend recent automated computer vision algorithms to reconstruct the global three-dimensional structures for photos and videos shot at fixed points in outdoor city environments. Mosaics of digital stills and embedded videos are georegistered by matching a few of their 2D features with 3D counterparts in aerial lidar imagery. Once image planes are aligned with world maps, abstract urban knowledge can propagate from the latter into the former. We project geotagged annotations from a 3D map into a 2D video stream and demonstrate their tracking buildings and streets in a clip with significant panning motion. We also present an interactive tool which enables users to select city features of interest in video frames and retrieve their geocoordinates and ranges. Implications of this work for future augmented reality systems based upon mobile smart phones are discussed.

**Keywords:** Computer vision, panorama, video, lidar, georegistration, geoquerying.

## 1. INTRODUCTION

The quantity, quality and availability of urban imagery have been rapidly increasing over the past few years. Billions of photos shot by inexpensive digital cameras within cities can now be accessed on the web. But retrieved urban thumbnails often bear little useful relationship with one another aside from having originated in some large, common metropolitan area. As anyone who has ever performed a Google image query knows, searching urban photo archives is currently a frustrating experience.

Fortunately, three-dimensional geometry provides an organizing principle for urban images collected at different times, disparate places and variable scales. Recall that photographs represent 2D angle-angle projections of 3D world-space subvolumes onto a variety of image planes. If the 3D structure of 2D image planes is reconstructed and georegistered with a map, high-level knowledge can propagate between the map and the pictures as well as among the images themselves. One may then interrogate urban photos for such basic information as the names of visible buildings, the distances to various movers, and the geolocations of interesting landmarks.

In this paper, we report upon a proof-of-concept demonstration of geoorganizing and geoquerying ground-level urban imagery. In our initial experiments, we focused upon the special but interesting case where all photos were shot from fixed world-space points. In particular, we collected digital stills and video clips at three different locations around MIT's campus where we allowed our cameras to rotate and zoom but not to translate. We sought to test whether abstract knowledge can be rapidly projected from 3D world-space into static panorama mosaics and subsequently into dynamic video sequences. As we shall see, this hypothetical question has an affirmative answer.

Our paper is organized as follows. In section 2, we first review an automated procedure for constructing 3D mosaics starting from a series of input photos. We then match a dynamic video stream with the static mosaic to provide human observers with useful spatial context. In section 3, we georegister a second mosaic and video stream with a laser radar (lidar) map. Geotagged annotations subsequently project from the 3D map into the video stream and track stationary

---

\* 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.

buildings and streets in a video clip with significant panning motion. In section 4, we embed a larger panorama mosaic and video stream inside a ladar map for all of Boston. A user can interactively query urban features within video frames

for their absolute geocoordinates and 3D ranges. Finally, we close in section 5 by summarizing our results and indicating directions for future work.

## 2. 3D MOSAICING OF URBAN PHOTOS AND VIDEOS

Video monitoring of urban environments has grown commonplace as camera technology increases in quality and decreases in price over time. Security cameras fixed atop poles or attached to building sides are routinely used to follow unusual movements within their fields of view. In the post 9/11 era, New York City is working towards establishing an entire network of linked ground cameras [1][2]. This video surveillance system will be patterned after London's "Ring of Steel" which directly led to the apprehension of the 2005 subway attack perpetrators. The objective for New York's ambitious camera network will be to monitor all vehicle and foot traffic within downtown Manhattan and uncover potential threatening activities.

Fixed security cameras can roam about their attachment points and zoom to provide soda-straw views of urban areas. However, they cannot simultaneously yield synoptic context which would help human eyes and brains better understand their instantaneous output. So it would be useful to embed a security camera's dynamic imagery inside a panoramic mosaic covering its accessible field of regard. We present such a computer vision capability within this section.

To begin, we review the basic procedure for generating mosaics from a set of input stills shot from a fixed location [3][4]. The first step is to automatically extract features from each input image based upon their intensity gradient contents. Over the past few years, the Scale Invariant Feature Transform (SIFT) has become a standard approach to photograph feature extraction which is relatively insensitive to varying camera perspective, zoom levels and illumination conditions [5]. SIFT yields a 128-dimensional vector descriptor for each image feature that captures local position, orientation and scale information. In order to rapidly compute SIFT features via graphics processor hardware, we employ the SiftGPU package [6]. Representative SIFT feature output is displayed for part of a 3264×2448 photograph in fig. 1.

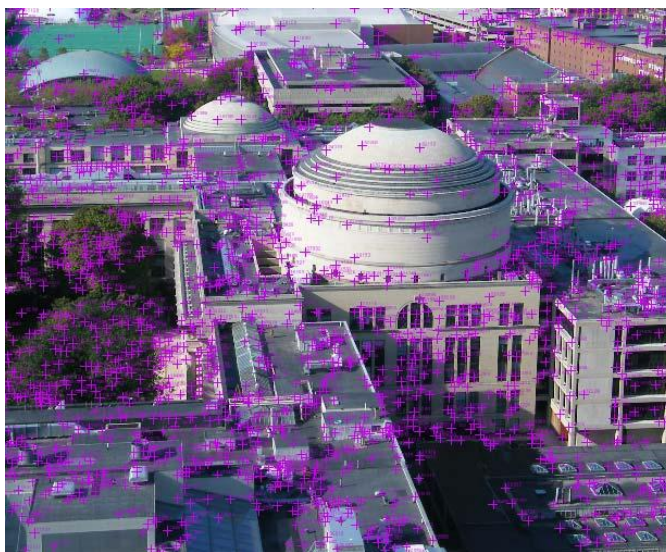


Fig 1: SIFT features within a photo shot from atop an MIT skyscraper.

Looping over pairs of input photos labeled by indices  $i$  and  $j$ , we next identify candidate bijective feature tiepoint matches via Lowe's ratio test [5]. Using Approximate Nearest Neighbor data structures with an  $\epsilon=3$  error bound to significantly decrease search times over the 128-dimensional vector space [7][8], we calculate distances  $d_1$  and  $d_2$

between the closest and next-to-closest candidate partner feature descriptors in photo  $j$  and an input descriptor in photo  $i$ . We accept the closest candidate feature as a genuine tiepoint match if  $d_1/d_2 < 0.5$ .

A number of incorrect feature matches slip through Lowe's ratio filter. So we employ an iterative Random Sample Consensus (RANSAC) algorithm to catch as many erroneous tiepoint pairings as possible [9]. Candidate matching features in photo  $i$  are partitioned into quadrants with respect to their median image plane location, and their bounding box is computed. Homographies are then constructed from four tiepoint pairs randomly pulled from different quadrants provided the spanning area of the points' quadrilateral exceeds 25% of the bounding box's area. Each four-point homography is next used to project all candidate tiepoint features in photo  $i$  onto photo  $j$ . If the distance between a projected feature and its candidate tiepoint counterpart is less than 0.005 where the photo's vertical axis is normalized to range between 0 and 1, we count the tiepoint pair as an inlier. The four-point homography maximizing the inlier count serves as the final classifier of tiepoint pair outliers. All surviving SIFT feature pairs are relabeled so that they carry a common ID. Results from this RANSAC matching procedure are displayed in fig. 2 for a few features in two overlapping photos. As can be seen in the figure, extracted tiepoint associations are generally reliable but incomplete.



Fig 2: Zoomed views of SIFT features found in overlapping regions of two different photos. Matched features share common ID labels.

Once matching features have been identified across all input photographs, our machine sequentially forms mosaics from subsets of the images ordered according to decreasing tiepoint pair count. Following Snavely *et al.* [10][11], we assume every photo's intrinsic camera calibration parameters are known except for a single focal length. The linear size of our Fujifilm S8000 camera's CCD chip along with its output EXIF metadata tags provide initial estimates for each photo's dimensionless focal parameter. Reasonable estimates for the camera's internal calibration matrices are then known, and 3D rays corresponding to matching 2D tiepoint features may be derived. Singular value decomposition of the matrix formed from summing outerproducts of corresponding 3D rays subsequently yields a rough guess for the relative rotation between images  $i$  and  $j$  [4][12].

Armed with initial estimates for all camera calibration parameters, we perform bundle adjustment via nonlinear Levenberg-Marquadt optimization using the LEVMAR package written in C [13]. We adopt the score function of Brown and Lowe which treats symmetrically every photo input to the panorama mosaic [3]. When iteratively solving for refined values of the focal and rotation parameters for the  $n^{\text{th}}$  image, we hold fixed the fitted camera parameters for the previous  $n-1$  images. After all photos have been added to the composite, we perform one final bundle adjustment where all camera focal and rotation parameters are allowed to vary.

Results from this mosaicing procedure for 21 photos shot during October 2008 from the rooftop of an MIT skyscraper are displayed in fig. 3. The relative orientation and solid angle size for each 2D image is depicted in the figure by its 3D view frustum. No attempt has been made to blend colors within the overlapping ensemble of photos. Nevertheless, the entire collection yields a high-fidelity, wide-angle view of a complex urban scene.

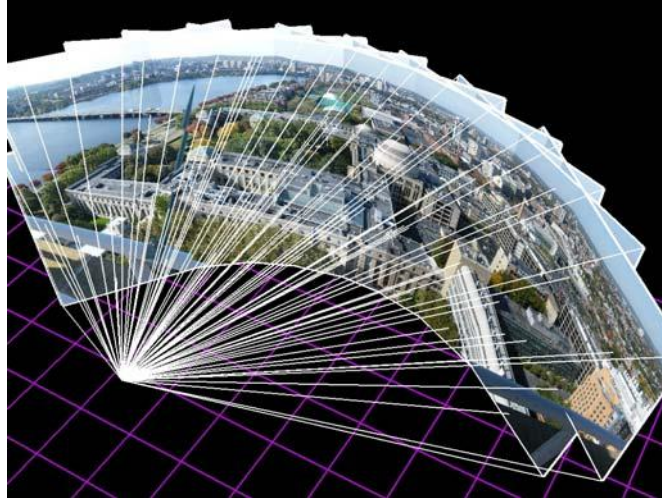


Fig 3: 3D mosaic of 21 photos shot from the MIT skyscraper.

After shooting the stills which comprise the panoramic mosaic, we replaced the digital camera on our skyscraper rooftop tripod with a high definition video camera. We subsequently collected footage inside the panorama's field of regard. For security camera output annotation purposes, we want to match each foreground video frame with the background mosaic as fast as possible. We therefore developed the following algorithm whose performance represents a compromise between accuracy and speed.

For each individual  $1888 \times 1062$  video frame, we extract SIFT features and match them with counterparts in the mosaiced photos which were precalculated once and stored. If a panoramic tiepoint partner is found for some video feature, a 3D ray is reconstructed from the calibrated still image and associated with the feature's 2D video coordinates. An iterative RANSAC procedure similar to the one employed for static panorama generation is utilized to minimize false correspondences between rays and coordinate pairs. A homogeneous matrix  $H$  mapping 3D world-space rays onto 2D video feature coordinates is subsequently calculated via least squares fitting. The entries in homography  $H$  may be directly transferred into the first  $3 \times 3$  block of a  $3 \times 4$  homogeneous matrix  $P$ . After taking the video camera's location as the world-space origin, we set the last  $3 \times 1$  column vector in projection  $P$  to zero without loss. All intrinsic and extrinsic calibration parameters for the video camera are thus recoverable for each video frame. This process independently matches each foreground video image to the background panorama. It runs at approximately 10 seconds per frame on a Dell Precision M6300 laptop.

Figure 4 displays the time-dependent view frustum for the video camera alongside the time-independent frusta for the mosaiced photos. In order to emphasize that the former is dynamic while the latter are static, we recolor the panorama pictures on a grey scale so that the colored video stream stands out more vividly. We also temporally smooth the  $3 \times 4$  projection matrices for every video frame via an alpha-beta-gamma filter [14][15]. The video stream then glides over the panorama with minimal jitter and also keeps up with sudden changes in camera pan and tilt. As the movie plays and roams around in angle space, it may be alpha-faded to reveal good agreement between the soda-straw and synoptic views.

The absolute position, orientation and scaling of the reconstructed frusta in fig. 4 cannot be determined by conventional computer vision techniques. To fix these global parameters, the photos and video stream must be inserted into a three-dimensional world map. We consider this problem in the following section.

### 3. GEOREGISTERING MOSAICS

The geometry of cities is highly complex. Fortunately, 3D urban structure can be efficiently measured by aerial laser radars. If the position of an airborne ladar is accurately known as a function of time and angle-angle-range information for each of its scans is recorded, XYZ coordinates for every surveyed ground point can be reconstructed on a

georegistered grid. Over the past decade, ladar technology has sufficiently matured so that sub-meter resolution maps for entire cities are now routinely generated in a matter of days.

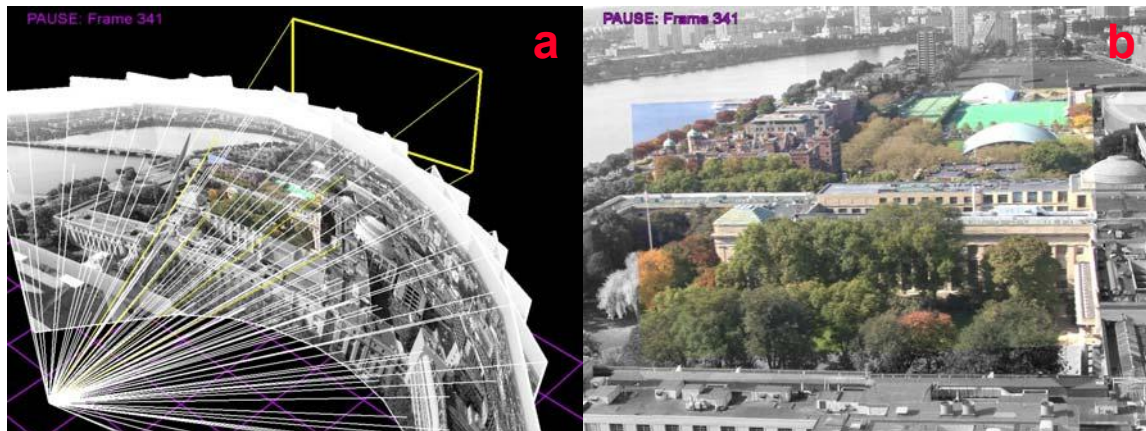


Fig 4: (a) Instantaneous angular location of one video frame relative to the background mosaic indicated by its yellow view frustum. (b) View of the dynamic video (colored) aligned with the static panorama (greyscale) from the ground camera's perspective.

Figure 5a illustrates an aerial ladar image covering a section of the MIT campus. The data were collected in 2007 by the JAUDIT system with a ground sampling distance of 25 cm. They are colored according to height via a color map designed to accentuate Z-content. Figure 5b exhibits a conventional aerial EO image covering the same general part of MIT which was snapped from Yahoo's website [16]. The 2D photo captures panchromatic reflectivities which the 3D ladar image lacks. So to maximize information content, we fuse the two together using an HSV coloring scheme [17]. The fused result is displayed on an absolute longitude-latitude grid in fig. 5c.

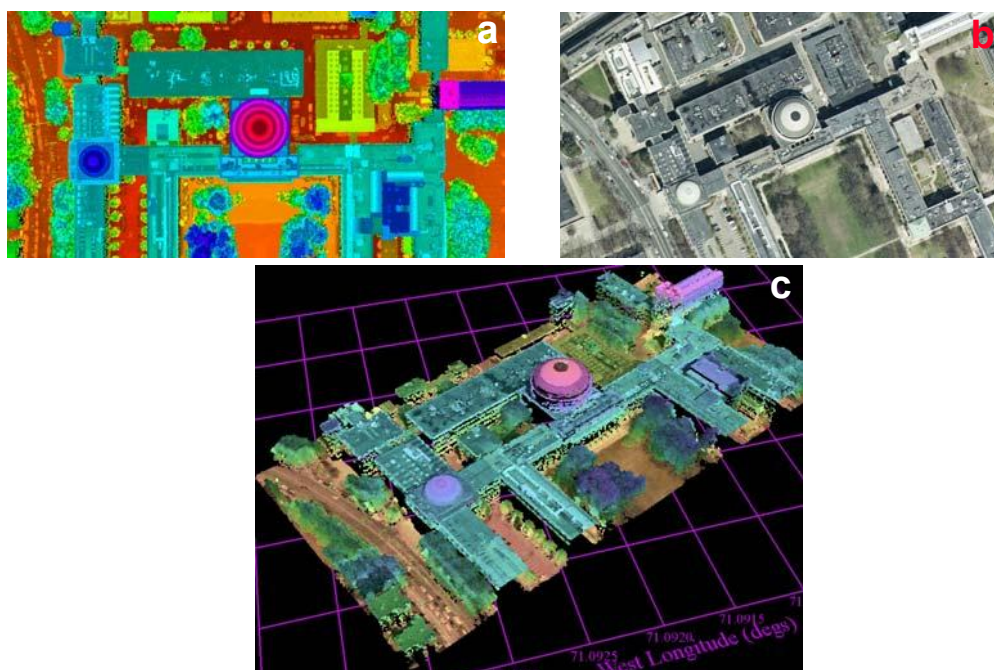


Fig 5: (a) Aerial ladar image of MIT campus colored according to height. (b) Aerial EO image of MIT campus naturally colored. (c) Fused 3D map.

In January 2009, we shot a second sequence of 14 ground photos from MIT's student union looking towards the university's main entrance which is located towards the lower left of fig. 5c. Following the same mosaicing procedure as for the first set of images collected from atop the skyscraper, we converted the overlapping 2D pictures into 3D view

frusta. Given the Yahoo aerial photo, it was relatively straightforward for us to determine the camera's absolute geolocation within the ladar map. But computing the global multiplicative factor by which each view frustum's focal length needed to be scaled was more involved.

We first manually selected 11 world-space points from the aerial ladar cloud which are visible in one or more of the 14 ground photos. We then quasi-randomly identified 22 two-dimensional partners for the 11 three-dimensional points in a subset of the calibrated photos. After converting the 2D features into averaged 3D rays, we compared the opening angle between each pair of backprojected panorama rays with the angle between corresponding world-space rays. The mean of world-space to image-space opening angle ratios yields the global multiplicative factor  $\rho$  by which the panorama must be rescaled. Equivalently, the focal parameter  $f$  for each still in the mosaic must be altered according to the formula

$$f_{new} = \frac{1}{2N_v \tan \left[ \rho \tan^{-1} \left( \frac{1}{2N_v f} \right) \right]} \quad (1)$$

where  $N_v$  denotes the number of vertical pixels in an image.

When we stretch the panorama by scale factor  $\rho$ , we must also adjust the relative rotation between each pair of mosaiced photos in order to maintain their alignment. This modification is readily derived if one works with angle-axis representations for  $3 \times 3$  rotation matrices. Each photo's new rotation with respect to some panorama origin point has the same axis  $\hat{n}$  as its original rotation. But its angle  $\gamma$  about  $\hat{n}$  must be rescaled by  $\rho$  when the panorama is stretched:

$$R_{new} = R(\hat{n}, \rho\gamma) \quad (2)$$

Application of scaling and relative rotation equations (1) and (2) renders the ground panorama's solid angle coverage consistent with the aerial ladar map.

In order to finish aligning the 2D photos with the 3D point cloud, we need to compute the global rotation  $R_{global}$  which transforms the rescaled bundle of image-space rays onto its world-space counterpart. So we again form a  $3 \times 3$  matrix sum of corresponding ray outerproducts and recover  $R_{global}$  from its singular value decomposition [4][12]. After the camera projection matrix for each mosaiced photo is rotated by  $R_{global}$ , we can insert the panorama into the ladar map (see fig. 6).

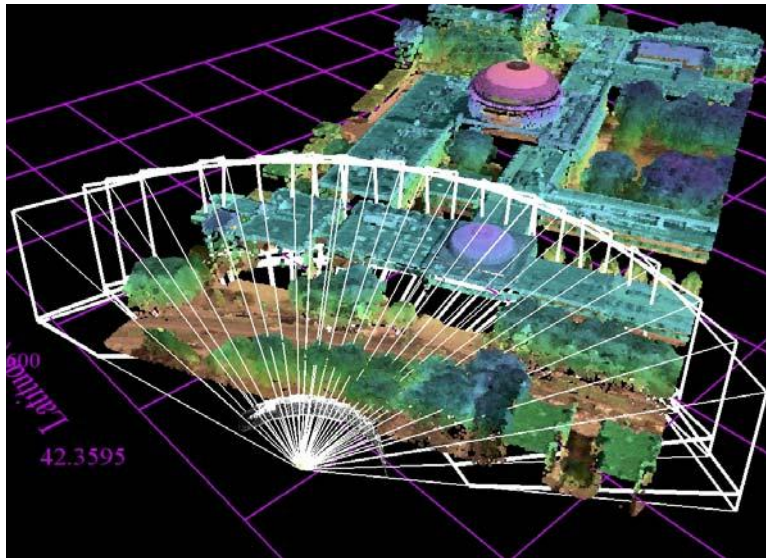


Fig 6: Ground-level panorama georegistered with the aerial ladar map.

Though the absolute position, orientation and scale of the photos' view frusta are fixed, the range at which we choose to view the image planes relative to the camera's location remains a free variable. By varying this radial parameter, we can visually inspect the alignment between the ground-level EO and aerial lidar data. In fig. 7a, the image planes form a ring relatively close to the ground camera and occlude most of its view of the lidar point cloud. If the ring's radius is increased as in fig. 7b, some lidar points emerge in front of the image planes. It is amusing to observe green leaves from the summertime lidar data "grow" on nude tree branches in the wintertime photos. The striking agreement between the tree and building contents in the 2D and 3D imagery confirms that the mosaic is well georegistered.

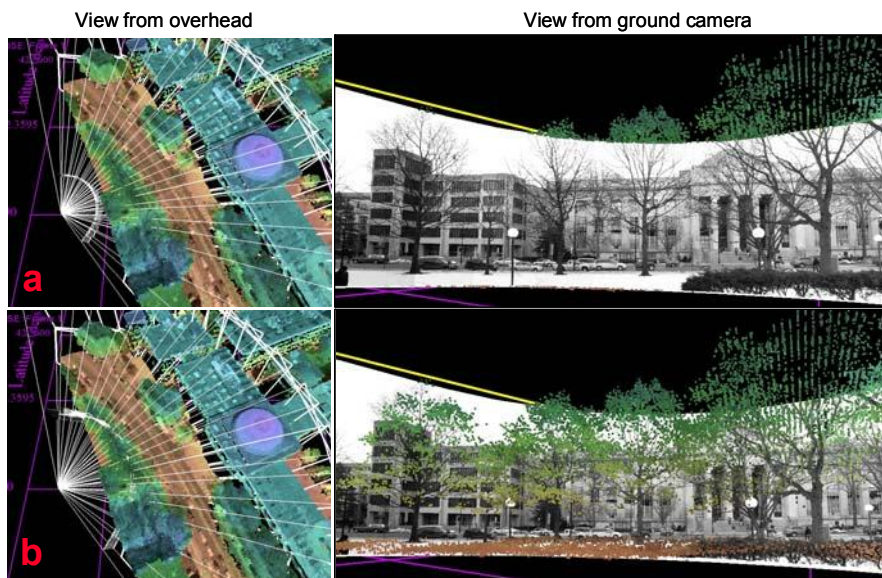


Fig 7: (a) Panorama photos occluding ground camera's view of lidar point cloud. (b) Summertime tree leaf lidar points appear to grow on nude wintertime branches as image planes' range is increased relative to ground camera.

Once the panorama is globally aligned with the point cloud, lidar voxels match onto corresponding photo pixels. Moreover, higher-level knowledge attached to the 3D voxels can propagate into the 2D image planes. Consider for instance names of buildings and streets, some of which are labeled for our MIT example in fig. 8a. Such urban information is frequently available through Geographic Information System (GIS) layers that enter into every standard mapping application currently running on the web. Building and street annotations have longitude, latitude and altitude geocoordinates which can project into calibrated photographs via their  $3 \times 4$  camera model matrices. Annotations identifying urban structures then automatically appear at correct locations within the static photos (see fig. 8b).

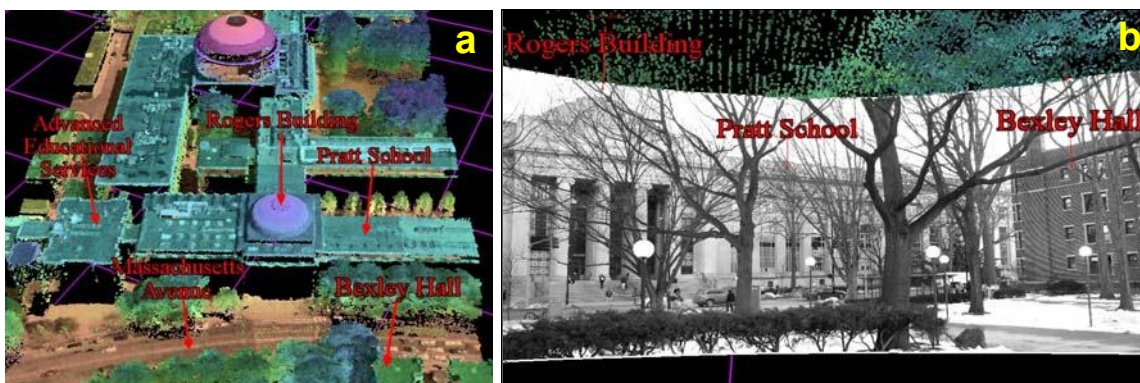


Fig 8: (a) Names of MIT buildings and streets appearing as geotags in 3D map. (b) Annotations automatically projected into georegistered mosaic.

Similar labeling of dynamic video clips shot in cities is possible provided they are georegistered with a 3D map. We follow the same matching procedure between our second stationary MIT background panorama and a colocated

foreground video as previously described for our first skyscraper example. The ground-level video sequence contains some foreground pedestrian and vehicle traffic which have no background counterparts in the mosaic. Nevertheless, our ray matching procedure successfully aligns the input video to the georegistered panorama with little difficulty (see fig. 9). Building and street names therefore project directly from world-space into the moving video stream (see fig. 10). As the movie plays, the annotations correctly track moving image plane locations for urban structures up to residual low-frequency jitter not completely removed by alpha-beta-gamma temporal filtering.

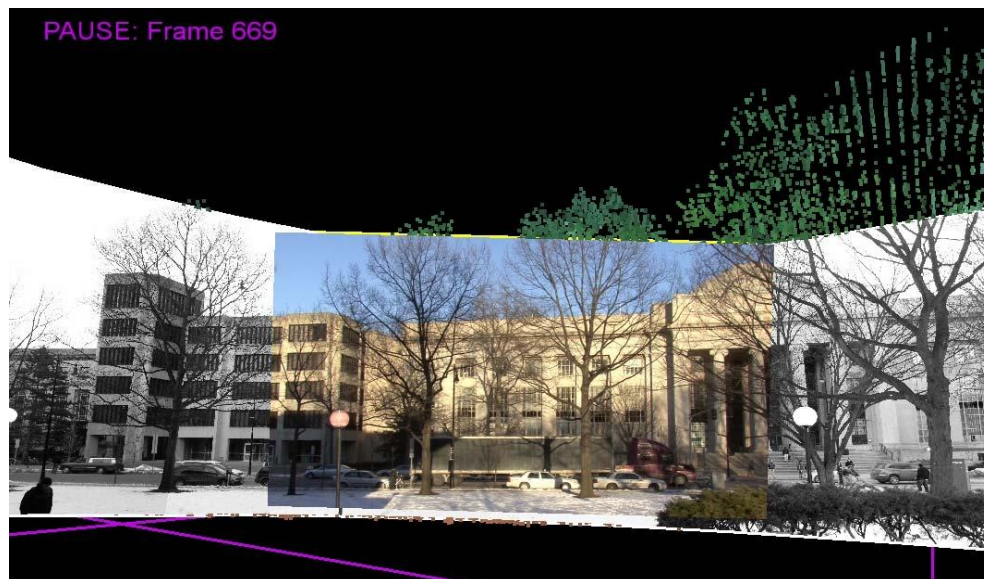


Fig 9: One frame from a video sequence automatically aligned with the georegistered mosaic. The dynamic and static imagery were both shot from the same ground location.

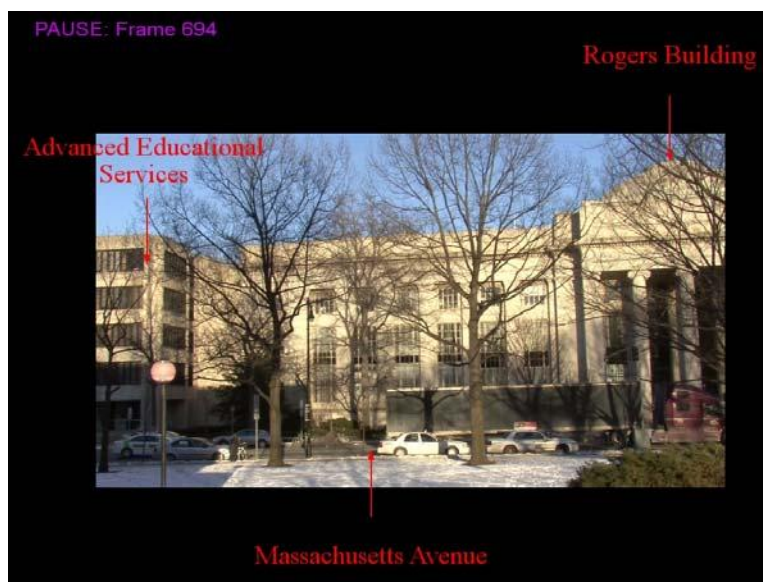


Fig 10: Annotation labels track stationary buildings and streets (and ignore moving vehicles) within a panning video camera clip.

This second MIT campus example proves that passive transfer of abstract information from 3D world-space into dynamic 2D image planes is possible. But it does not yet demonstrate active image-based querying of urban knowledge. We introduce such an interrogation capability in the next section.

#### 4. GEOQUERYING URBAN VIDEOS CLIPS

Estimating absolute distances within conventional photographs is generally difficult. For one or more photographs shot from a fixed location, it is mathematically impossible to derive any range information via triangulation. For two or more images collected by cameras with nonzero baseline separation, stereo reconstruction of common features can yield relative distances but with errors that increase as angular separations decrease. Perspective effects render urban scene geoinformation extraction particularly confusing. When looking at city structures, one often cannot tell whether some building is taller than another or closer in space. In order to accurately perform such mensuration, an image analyst needs genuine three-dimensional input.

Interrogating 2D images for absolute heights and ranges becomes tractable if they are embedded in a 3D map. Figs. 11a and 11b illustrate a still panorama of the Boston skyline constructed from 73 input photos along with a dynamic video sequence inside a large aerial lidar point cloud for the entire city. The photos and video were shot in 2009 from atop MIT's tallest skyscraper looking over the Charles River towards Boston's financial and commercial centers. On the other hand, the Topographic Engineering Center lidar data were gathered in 2004 with a 1 meter ground sampling distance. Given the 5 year gap between the 2D and 3D data collections, some new skyscrapers present in the former do not exist in the latter. But otherwise, the match between the EO and lidar imagery looks qualitatively good (see fig. 12).

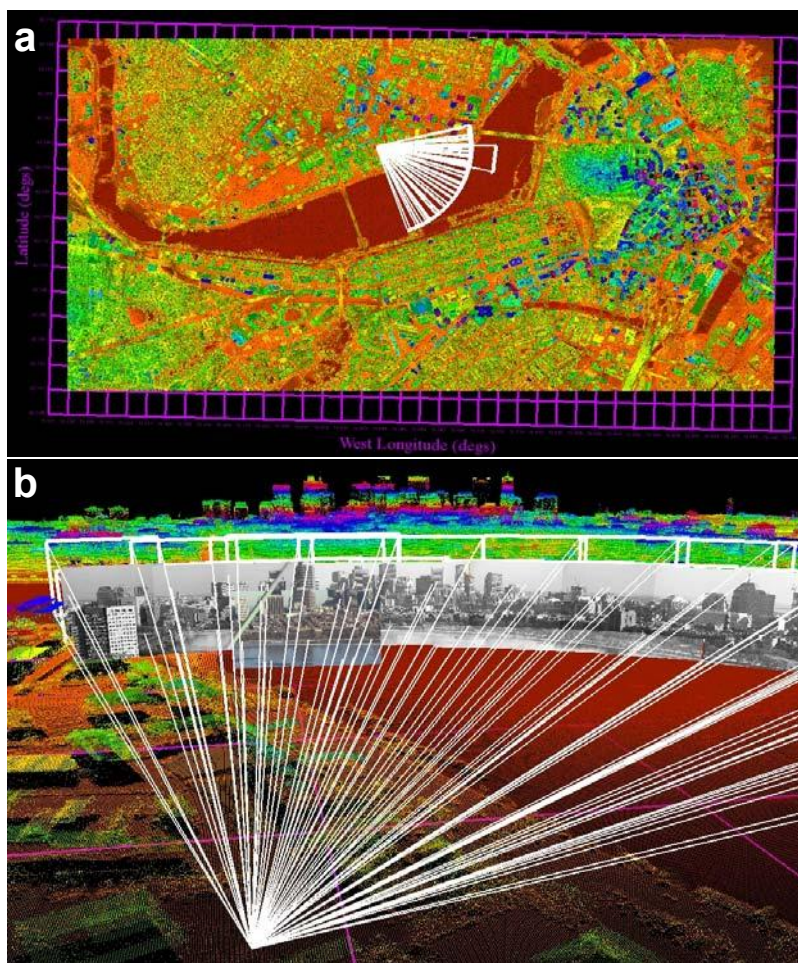


Fig 11: (a) Overhead and (b) ground-level views of Boston skyline panorama and video clip georegistered with aerial lidar map.

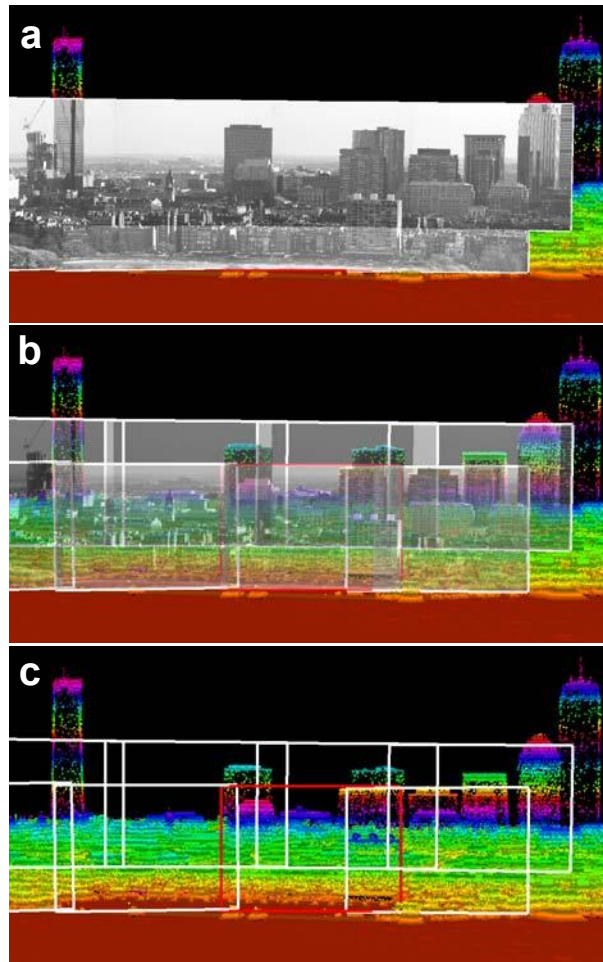


Fig 12: Comparison of Boston panorama and ladar point cloud from ground camera's perspective with (a) 0%, (b) 50% and (c) 100% alpha blending.

To exploit this geoalignment, we have developed a prototype image-based querying tool which plays the video in one window and superposes the movie's frustum on the Boston point cloud in another window. When a user selects a 2D video pixel, a corresponding 3D voxel is calculated via ray tracing. The point in the ladar cloud is selected based upon its angle relative to the camera's line-of-sight as well as its range relative to the camera's location. The 3D counterpart is marked by a blinking set of crosshairs. It is also projected back into the 2D image planes for all dynamic frames in which it was visible to the video camera, including the original frame used to select the feature's initial pixel coordinates. This reprojection guarantees a valid mathematical relationship between the feature's 3D and 2D tiepoint coordinates.

Once a counterpart voxel to a queried pixel is found, the former's geocoordinates automatically transfer to the latter. In approximately one second on our Dell M6300 laptop, the prototype software tool returns and displays the selected video feature's longitude and latitude (see fig. 13). Alternatively, our interactive application can report feature range from the camera as well as altitude above sea-level (see fig. 14). Such geoinformation is propagated to all other video frames in which selected urban features are visible. With our tool, a user can discover the heights for tall skyscrapers, the distances between various city points, and the geopositions for moving cars of interest. It consequently exhibits a nontrivial capability to query urban digital imagery for abstract knowledge.

## 5. SUMMARY AND FUTURE WORK

In this proof-of-concept paper, we have demonstrated how 2D photos and videos shot from fixed locations within a city may be inserted into 3D world-space to enable automatic propagation of urban knowledge between maps and digital

imagery. Georegistration of photo mosaics with ladar data requires a one-time manual selection of a few dozen tiepoint pairs. But all other components of our processing pipeline are fully automated. Building and street name geotags initially project from ladar maps onto panorama stills. After thousands of video frames are subsequently matched to panoramas, urban structure annotations transfer from static backgrounds onto dynamic foregrounds. Moreover, users may select stationary targets of interest within video frames and retrieve their corresponding geocoordinates and ranges.

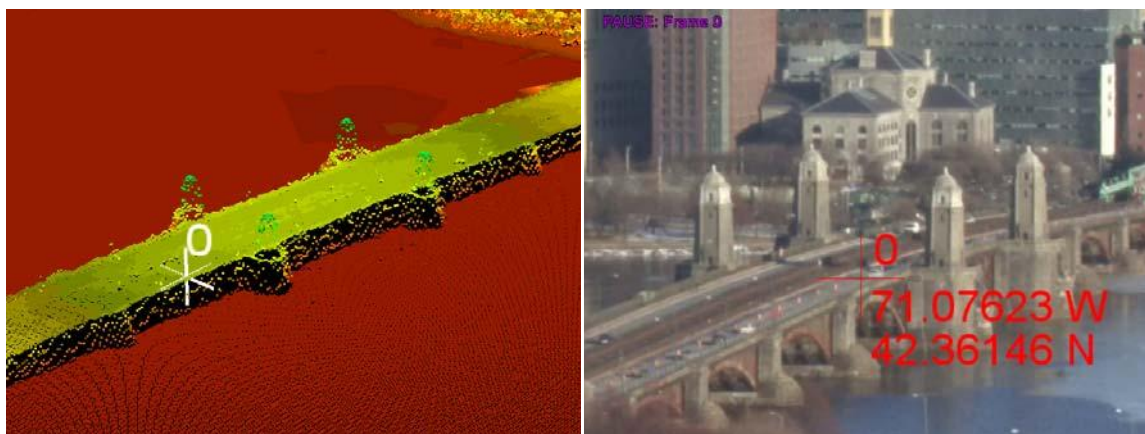


Fig 13: A Boston bridge pixel selected from a video frame is automatically raytraced to a corresponding voxel in the ladar point cloud. The queried pixel's longitude and latitude are returned and displayed in the video window.

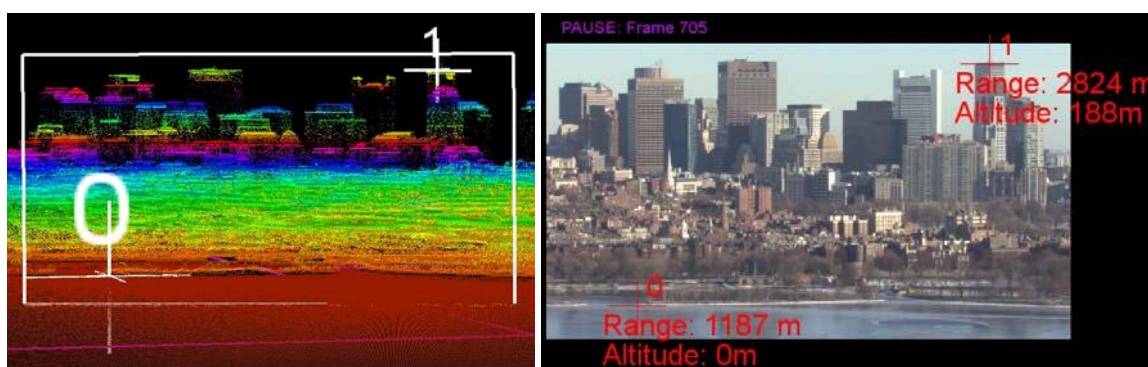


Fig 14: Selected sea-level and skyscraper rooftop pixels are automatically raytraced to corresponding voxels. (The video camera's instantaneous field of view is depicted by the white rectangle in the ladar window.) The queried pixels' ranges and altitudes are displayed in the video window.

The basic idea of using 3D world geometry to organize 2D imagery should be extended in the near future to sets of urban photos and videos more general than those considered here in our initial experiments. By combining the recent pioneering "Photosynth" computer vision work of Snavely *et al.* [10][11] with wide-area ladar maps for entire cities, it should be possible to georegister thousands of outdoor city photos in a similar fashion as for the smaller numbers of mosaiced images in this work. Once a sufficiently large archived set of 2D photos has been reconstructed and georegistered, new digital pictures gathered by mobile devices such as smart phones could rapidly be incorporated into the 3D map just as our video sequences were matched with the panorama stills. Near real-time annotation of such live photo inputs would endow smart phones with augmented urban reality capabilities. Given the dramatic rise in the quantity of digital images over the past few years, we believe the impact of systematically organizing and exploiting vast numbers of photos could someday become comparable to that of Google text search.

We therefore look forward to extending the preliminary results reported here and pursuing this exciting potential for image-based querying of urban knowledge databases.

## REFERENCES

- [1] J. Mullins, *Ring of steel II - New York City gets set to replicate London's high-security zone*, IEEE Spectrum, **43** (2006) 12.
- [2] C. Buckley, *Police Plan Web of Surveillance for Downtown*, The New York Times, July 9, 2007.
- [3] M. Brown and D.G. Lowe, *Automatic Panoramic Image Stitching using Invariant Features*, International Journal of Computer Vision, **74** (2007) 59.
- [4] M. Brown, R.L. Hartley and D. Nister, *Minimal Solutions for Panoramic Stitching*, IEEE Conf. on Computer Vision and Pattern Recognition (2007) 1.
- [5] D.G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision (2004) 91.
- [6] C.C. Wu, *A GPU implementation of David Lowe's Scale Invariant Feature Transform*, (2007), downloadable from <http://cs.unc.edu/~ccwu>.
- [7] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman and A.Y. Wu, *An Optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions*, Journal of the ACM (1998) 891.
- [8] D.M. Mount, *ANN Programming Manual*, (2006), downloadable from <http://www.cs.umd.edu/~mount/ANN>.
- [9] M. Fischler and R. Bolles, *Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography*, Communications of the ACM (1981) 381.
- [10] N. Snavely, S.M. Seitz and R. Szeliski, *Photo Tourism: Exploring Photo Collections in 3D*, ACM Transactions on Graphics, **25** (2006) 835.
- [11] N. Snavely, S.M. Seitz and R. Szeliski, *Modeling the World from Internet Photo Collections*, International Journal of Computer Vision, **80** (2008) 189.
- [12] K.S. Arun, T.S. Huang and S.D. Blostein, *Least-Squares Fitting of Two 3-D Point Sets*, IEEE Transactions on Pattern Analysis and Machine Intelligence (1987) 698.
- [13] M. Lourakis, *LEVMar* (2008), downloadable from <http://www.ics.forth.gr/lourakis/levmar/>.
- [14] P. R. Kalata, *The Tracking Index: A Generalized Parameter for  $\alpha$ - $\beta$  and  $\alpha$ - $\beta$ - $\gamma$  Target Trackers*, IEEE Transactions on Aerospace and Electronic Systems (1984) 174.
- [15] J.E. Grays, *A Derivation of an Analytic Expression for the Tracking Index for the Alpha-Beta-Gamma Filter*, IEEE Transactions on Aerospace and Electronic Systems (1993) 1064.
- [16] See <http://maps.yahoo.com> for aerial urban imagery.
- [17] P. Cho, *3D Organization of 2D Urban Imagery*, Proc. Of SPIE: Signal Processing, Sensor Fusion and Target Recognition XVII, **6968** (2008).