International Journal of Computer Vision, 53(1), pp. 93--107, June 2003

Calibrated, Registered Images of an Extended Urban Area

Seth Teller Matthew Antone Zachary Bodnar
Michael Bosse Satyan Coorg Manish Jethwa Neel Master
MIT Computer Graphics Group
Email: {teller,antone,zbodnar,ifni,satyan,manish,neel}@graphics.lcs.mit.edu

Abstract

We describe a dataset of several thousand calibrated, time-stamped, geo-referenced, high dynamic range color images, acquired under uncontrolled, variable illumination conditions in an outdoor region spanning several hundred meters. The image data is grouped into several regions which have little mutual inter-visibility. For each group, the calibration data is globally consistent on average to roughly five centimeters and 0.1°, or about four pixels of epipolar registration. All image, feature and calibration data is available for interactive inspection and downloading at `http://city.lcs.mit.edu/data`.

Calibrated imagery is of fundamental interest in a variety of applications. We have made this data available in the belief that researchers in computer graphics, computer vision, photogrammetry and digital cartography will find it of value as a test set for their own image registration algorithms, as a calibrated image set for applications such as image-based rendering, metric 3D reconstruction, and appearance recovery, and as input for existing GIS applications.

1 Introduction

This paper describes data produced by a system for calibrated, terrestrial image acquisition in urban areas. The system includes a novel sensor, and a suite of scalable geometric algorithms, which produce accurately calibrated, geo-referenced terrestrial (near-ground) imagery of urban scenes with no human intervention or interaction required. The system is end-to-end, in the sense that it acquires uncalibrated images as input, and produces geo-referenced CAD models as output, with no human interaction other than the deployment of the sensor. Detailed descriptions of the system's design rationale, components, and algorithms appear elsewhere [Tel97, Tel98, CMT98, BdT99, AT00, AT01, AT02]. This paper describes an extensive collection of calibrated image data produced by the system [TAB$^+$01], which we have placed on-line for interactive viewing and download at `http://city.lcs.mit.edu/data`. To our knowledge, this dataset represents the largest collection of calibrated terrestrial imagery in existence.

We envision at least three ways in which the data may be useful to others. First, the uncalibrated imagery (i.e., data from early stages of our processing pipeline) can be used as test data by researchers developing large-scale image calibration and registration algorithms. Second, the registered imagery (i.e., data from the end of the process) can be used "as is" for a variety of applications including image-based rendering and 3D reconstruction. In either context, the scale and extent of the data we present significantly exceeds that of any existing real data set; thus its availability should pose an interesting collection of challenges. Finally, we note that since the data is expressed in a geo-referenced (Earth) coordinate system, it can be readily incorporated into a variety of existing GIS and digital cartography applications (e.g. OpenGIS [Ope], TerraServer [Ter], and the National Spatial Data Infrastructure [NSD]).

The paper is organized as follows. Section 2 describes the acquisition and processing stages in our system. Section 3 describes a collection of objective performance (accuracy, consistency) measures for our methods, and the results of applying these measures to our data. Section 4 describes the web interface to the dataset. Section 5 describes existing acquisition methods for geo-referenced imagery. Section 6 summarizes the contributions of the paper, and an Appendix details the data formats and conventions used for representing image, calibration, and feature data.

2 Calibration Stages

The sensor is deployed in acquisition "sessions." After data upload, a series of calibration and processing stages revises image data or metadata as follows:

1. Off-line (semi-automated) intrinsic camera calibration;
2. Off-line (semi-automated) photometric camera calibration;
3. Off-line design of "tiling" for omni-directional image mosaics;
4. Acquisition of HDR imagery, with approximate geo-referenced pose for each image;
5. Data upload, spatial indexing, and generation of node adjacency graph;
6. Radial distortion correction;
7. Image pyramid generation;
8. Mosaic generation and refined intrinsic calibration;
9. Sub-pixel edge and point feature detection;
10. Rotational registration (registration to scene vanishing points);
11. Translational registration (up to absolute scale, offset, and rotation); and
12. Final registration to geo-referenced (i.e., Earth-relative) coordinates.

2.1 Off-Line Intrinsic Calibration

Variations in lens attachment, temperature, etc. can perturb intrinsic parameters. At the start of each acquisition session, the sensor acquires several images of a calibration pattern in order to recover initial estimates of the camera intrinsics and the radial lens distortion parameters. We use a calibration pattern constructed in such a way that 221 calibration points can be detected automatically and localized to sub-pixel accuracy (Figure 1). The coordinates of the calibration marks are then processed with a public-domain implementation of Zhang's calibration algorithm [Zha98], which numerically optimizes the camera's intrinsic parameters: focal length, aspect ratio, skew, principal point, and the first- and second-order radial distortion parameters. Later, our algorithms use the intrinsic parameters to remove radial distortion through resampling (Figure 2).

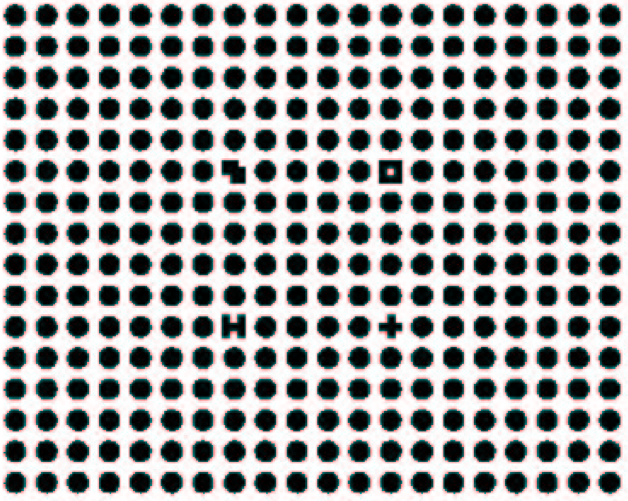


Figure 1: The pattern used for initial intrinsic calibration.

2.2 Off-Line Radiometric Calibration

The amount of light entering the camera varies as the square of the aperture diameter. We store the camera aperture value during acquisition, and later adjust the acquired pixel values to account for aperture variation across nodes (Figure 3). This allows comparison and combination of image pixels acquired with different apertures (i.e., under different lighting conditions).

Absolute radiometric calibration need be done only once for a particular camera CCD. For calibration we expose the camera to a bright indoor light source (not the sun) and acquire a high-dynamic-range (HDR) image [DM97], then set an absolute radiance scale such that the brightest pixel values map to 1.0 (i.e., zero on a log scale). Dark image calibration is not necessary since the magnitude of heat noise at our longest exposure time (1/10th second) is insignificant. A pixel value of zero is under-saturated and does not have a valid radiance value; such pixels are ignored in further processing. In practice only very few such pixels are present in our data, since all images are acquired during the day under adequate lighting.

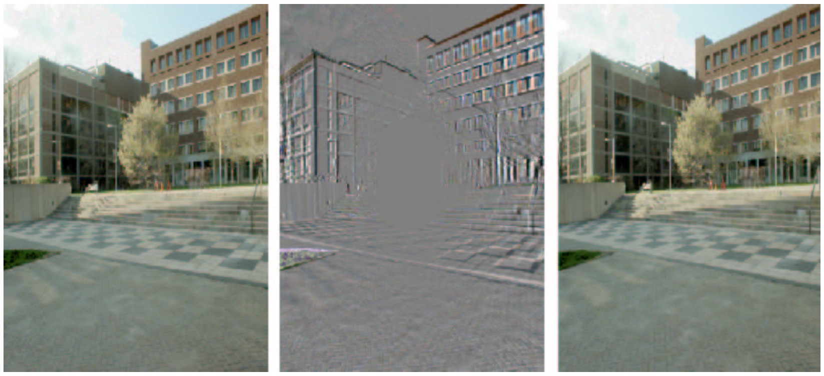


Figure 2: Left: Raw image. Right: Radial distortion removed. Center: Difference image.

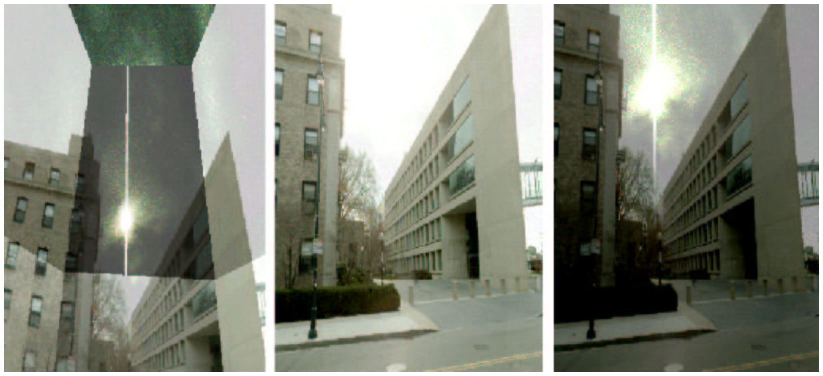


Figure 3: Left: Raw log-radiance images, each with distinct radiance ranges. Center: Images using a common radiance range. Right: the same images, rendered with a different radiance scale.

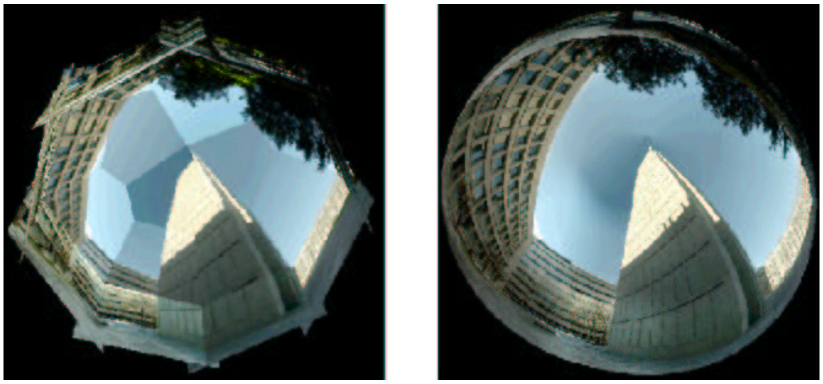


Figure 4: Left: A tiling of 20 images. Right: After mosaic generation.

2.3 Mosaic Design

Once reasonably accurate camera calibration is available, we have enough information to design the mosaic, or tiling pattern, which will cover a portion of the sphere during omni-directional image acquisition (Figure 4). Designing the tiling pattern requires only a rough FOV estimate [CT00]. Starting with this estimate we choose an overlap percentage, usually about 15%, and generate a series of azimuthal and altitudinal camera rotations that tile the sphere while preserving the desired degree of overlap. The tiling generation program also produces the adjacency edges for the tiling, that is, the set of image pairs with significant overlap.

2.4 High-Dynamic-Range Image Capture

The camera is mounted with its fixed optical center centered at the center of rotation of an electromechanically actuated pan-tilt head [DeC98]. At each head position, the camera captures one tile of the image mosaic at several exposures, averaging multiple frames at each exposure time to reduce image noise. The averaged images are then combined to produce a high-dynamic-range (HDR) image [DM97]. This enables pixels that were saturated in one frame to be replaced by pixels arising from a shorter exposure time.

Conventional 8-bit linear pixel encodings are not sufficient to store HDR imagery. Hence the images are initially stored using a 16-bit logarithmic encoding for each RGB value. Later in the pipeline, the images are converted to SGI-format `.rgb` files, with each pixel value storing a log-radiance value from the HDR imagery. This format enables processing and viewing by conventional tools. Since the original 16-bit image is preserved, images can also be converted to other formats for richer representation of dynamic range, such as LogLuv format [Lar98].

The sensor annotates each acquired HDR image with a camera descriptor, date- and time-stamp, camera intrinsics and estimated Earth-relative position and orientation [BdT99, DeC98]. The sensor's raw pose estimates are typically accurate to a few meters of position and a few degrees of orientation, but can be worse if GPS conditions are particularly poor during acquisition, for example due to satellite obscuration, multi-path reception or electromagnetic interference.

2.5 Data Upload, Spatial Indexing, and Adjacency Graph

After each acquisition session, the sensor rig is returned to the lab and reconnected to the local network. Its acquired data is then uploaded to the project's computational servers. Upon upload, each node is inserted into an abstract spatial index [PS85] keyed on absolute camera position (which may be revised by subsequent extrinsic calibration stages). This enables efficient computation of the dataset "adjacency graph", a list of each node's k nearest neighbors (we typically use $k \leq 6$), as well as inverse-range queries (e.g., "which nodes fall within the specified region?").

2.6 Correction of Radial Distortion

Using intrinsic parameters recovered earlier, the images are resampled to remove radial lens distortion (Figure 2), enabling downstream computations to use a simple pinhole camera model. The images are clipped to have approximately the same size and central resolution as the original images.

2.7 Image Pyramid Generation

After the full-resolution images have been undistorted, they are filtered down to half-, quarter-, and $3/32-$resolution to form a four-level gaussian image pyramid [Ros84]. Full-resolution images are used for feature extraction (see below). Half- and quarter-resolution images are used for multi-resolution mosaic generation, 3D reconstruction, and texture estimation [CT99, WTT$^+$02]. The thumbnail (3/32-resolution) images are used only for fast visualization.

2.8 Mosaic Generation

Our extrinsic calibration algorithms treat each node as a rigid, effectively wide field of view image, drastically reducing the number of extrinsic DOFs to be recovered per node. However, the sensor's raw rotation estimates are not sufficiently accurate to combine the image tiles directly. Thus, we use a correlation-based optimization algorithm [CT00] to estimate the rigid camera rotations relating the tiles (cf. Figure 4). This mosaic generation stage takes the acquisition rig's rotation estimates as inputs, then recovers improved rotations for each level of the image pyramid, using each level's converged estimates as initialization for the mosaic of the next higher resolution. The mosaic algorithm also refines the system's estimates of the camera's intrinsic parameters.

The "spherical images" produced by the mosaic stage are used only for visualization; whenever an image sample or feature is needed by any batch processing stage, the system samples directly from the raw, conventional images (using the per-image rotation estimates produced by mosaic generation). This avoids resampling the source images, and the attendant loss of information that would cause. For convenience, we also produce a six-sided cubical "environment map" of the spherical field-of-view; this too is used only for visualization in our system but could be used by others in different ways.

2.9 Sub-Pixel Edge and Point Feature Detection

Our registration algorithms do not use image pixels directly, but rather use edge and point features (Figure 5). Linear (edge) features are extracted through a two step process: first, sub-pixel zero crossing contours of the Laplacian of the Gaussian of the image are found. Then the edge contours are recursively split and fitted onto straight line segments, which we adopt as edge features.

Point (corner) features are generated by intersecting edges that lie sufficiently close together in image space, and form a large enough angle, to plausibly arise from a building, window, or other real-world corner. Our system forms intersections from edge features that are separated by at most 2° in image space (or about 40 pixels, at our highest image resolution, 1 milliradian per pixel), and that form an angle in image space of at least 5°.

2.10 Rotational Registration

The rotational registration stage, described elsewhere [AT00, AT02], takes intrinsic calibration information, edge features, and the node adjacency graph as inputs. It groups observed edge features into scene-relative vanishing points (VPs). Each node is assumed to have viewed a set of VPs that overlaps with or is identical to the set of VPs observed by its neighbors; nodes are brought into rotational alignment by registering each to the set of commonly observed VPs in its vicinity. This method brings the nodes into rotational alignment to within about 0.1°, or roughly

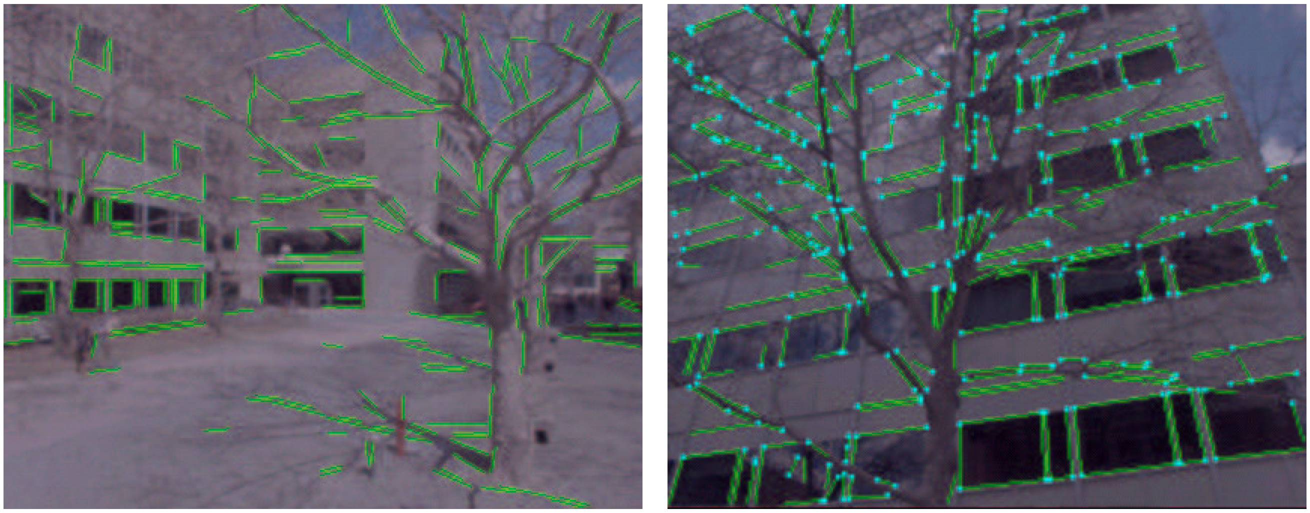


Figure 5: Edge (left) and point (right) features for two images in the dataset.

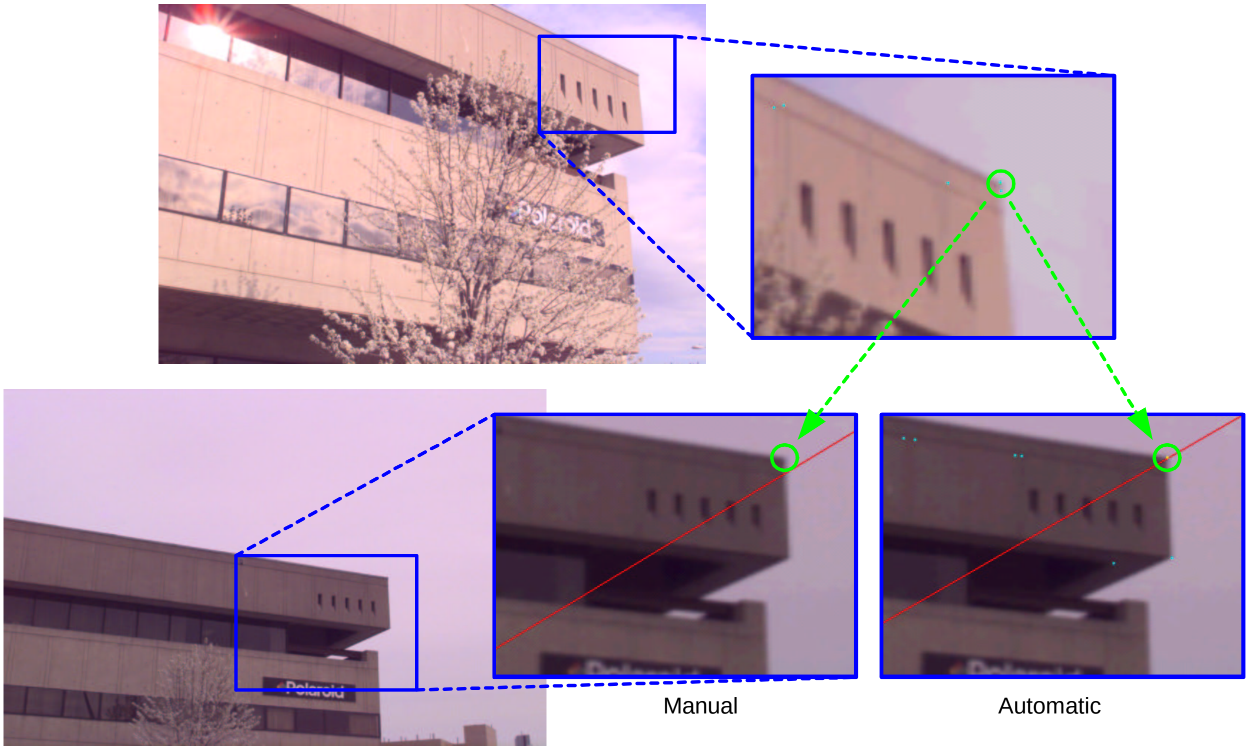


Figure 6: Epipolar registration resulting from manual bundle-adjustment [CT00] (bottom middle), and from our automated algorithm [AT02] (bottom right).

two pixels at our sensor resolution. (That is, if the point at infinity corresponding to a single VP direction is projected into multiple images, its image coordinates will be uncertain by about two pixels.) Section 3 describes consistency metrics for rotational registration.

2.11 Translational Registration

The translational registration stage, described elsewhere [AT01, AT02], takes intrinsic calibration information, point features, rotation information from the previous stage, and the node adjacency graph as inputs. It produces revised position estimates for every node, subject to pairwise baseline directions determined for each node adjacency. (This stage also produces, as a side effect, a set of probabilistic correspondences between point features across all pairs of adjacent nodes.) The resulting pose assignment is valid up to an arbitrary Euclidean transformation (translation, rotation and isotropic scaling). The quality of the epipolar geometry (Figure 6) can be assessed with a variety of consistency metrics (Section 3).

2.12 Geo-Referencing

The final processing stage registers the node set to the original GPS (i.e., absolute, geo-referenced) position estimates, exercising the final rigid translation, scaling, and rotational degrees of freedom for the entire dataset [AT01, AT02]. The resulting pose estimates are geo-referenced to an absolute (Earth) coordinate system and are metrically meaningful. The coordinates are stored, in units of meters, relative to a local tangent plane with its origin defined as the location of our GPS base station.

The Cartesian ECEF (Earth Centered Earth Fixed) coordinate system is based on the WGS-84 Geoid [HWLC97], the native geoid for GPS. The Euclidean transformation between Cartesian ECEF coordinates and the Cartesian LTP (Local Tangent Plane) coordinates is well-defined. We reexpress global coordinates with respect to the local tangent plane because LTP coordinates are more convenient and more accurate in floating point computations: the LTP z axis is nearly aligned with the local vertical, and LTP node coordinates have small magnitude.

3 Consistency Metrics

There is no ground truth for this dataset. We estimate that either manual surveying or semi-automated bundle adjustment on the dataset would require tens or hundreds of person-hours. Moreover, the dataset includes observations of many scene features that are impossible to localize directly, and safely, due to their physical placement in the world – for example, building corners high above ground. Therefore we have formulated a set of objective consistency metrics for the dataset, each of which assesses the degree to which separate observations are mutually consistent. Our consistency metrics take into account both local and global properties of the dataset.

The remainder of this section describes the result of applying a variety of consistency metrics to three calibrated image sets acquired at three regions on campus with limited intervisibility [AT02]. The datasets include between 1,000 and 4,000 images, spanning areas ranging from 100×100 meters to 300×400 meters, with camera altitude varying by only a few meters.

In addition to the automatically evaluated metrics defined below, the web interface to the data (Section 4) provides interactive tools for qualitative inspection of raw images, mosaic quality, extracted features, epipolar geometry, etc.

3.1 Local Consistency Metrics

We evaluated two local consistency metrics for each mosaic, that is, for each collection of images acquired at a single optical center. The local consistency metrics are defined as follows:

- **Mosaic pairwise cross-correlations.** For each node, we computed the sum, over all adjacent images within the mosaic [CT00], of the pixel inverse cross-correlation normalized by the number of overlapping pixels and the pixel dynamic range. A perfectly registered mosaic, assuming constant illumination and no resampling errors, would exhibit an inverse cross-correlation of 0.0, the minimum. A mosaic in which black pixels everywhere overlap white pixels would exhibit an inverse cross-correlation of 1.0, the maximum. The cross-correlation of each mosaic process at termination, typically between 0.005 to 0.05, is stored as a `RESIDUE` value in the pose descriptor file associated with the node (Section A.3).

- **Node VP spreads.** When acquired in an urban area, each node typically observes one or more vanishing points, or families of parallel lines in the scene. For truly parallel world lines, perfectly calibrated mosaics, and error-free edge features, the vanishing point could be estimated with no error. In reality, all of these elements exhibit noise, so we represent vanishing points as projective probability densities [Bin74, AT00]. We evaluated the width of the density, in degrees, at 95% confidence – that is, the size of the symmetric region of the density which includes 95% of its probability. In our data, each node observed 3-4 VPs on average; vanishing points within each node were estimated to within about 0.1° (2 milliradians).

3.2 Global Consistency Metrics

Our global consistency metrics are those that apply to collections of nodes at different positions in space, and to the scene quantities (vanishing points and point features) to which the mosaics are registered. We use these global consistency metrics:

- **Scene VP spreads.** When multiple mosaics observe a single vanishing point, the degree to which they agree on the absolute direction of the VP is a measure of consistency. For each globally observed VP, we determined the number of nodes that observed it, and the width of the consensus VP direction (using the 95% confidence bound described above). Typical global VP variances were 1-2 milliradians $(0.05 - 0.1^\circ)$ on average, and 2-4 milliradians $(0.1 - 0.2^\circ)$ worst case, or about 1-4 pixels of misalignment at our highest image resolution [AT00].

- **Node position spreads.** We computed uncertainty estimates for the recovered node positions by evaluating the average and maximum sizes at which 95% confidence bounds are reached for the recovered Gaussian densities. Node positions were estimated to 5-6 centimeters on average, and 8-11 centimeters worst case [AT02].

- **Epipolar residuals**. Our translation registration algorithm produces soft (probabilistic) correspondences between observed scene points. For each soft point match with probability greater than a threshold (we use $p \geq 0.8$), we evaluated the mean, maximum, and variance of the distance in image space between the 2D point feature and the epipolar line of its correspondent feature [AT02]. Epipolar alignment was consistent on average to within 1-4 pixels, with worst-case deviation 5-6 pixels and standard deviation about 2 pixels.

- **Pairwise 3D Point feature residuals.** We assessed end-to-end 3D feature consistency using the thresholded match probabilities, by extruding the implicated point features to 3D rays, and evaluating the average and maximum 3-D distance (in centimeters) between rays extruded from adjacent nodes. The mean and maximum residuals were 10-15 and 13-20 centimeters, respectively, with a standard deviation of 3-6 centimeters [AT02].

4 Web Interface

An interactive interface to the dataset is available at `http://city.lcs.mit.edu/data`. The interface depicts acquired nodes overlaid on a geo-referenced map (Figure 7). Each node is color-coded by the type of calibration metadata available for the node (some nodes have no revised position estimates, so are posted only with orientation estimates). The user may select any individual node for examination, producing a node inspection page in which the full node mosaic and the node's constituent (log-radiance) images are displayed (Figure 8). This page also displays the omni-directional image mosaic for the node, which can be panned and zoomed interactively, exposure-adjusted, and overlaid with extracted features and other information. Each node page includes links to the node's raw image data, ASCII intrinsic and extrinsic calibration data, and consistency information. Near the mosaic viewer is a "mini-map" of node context, showing the node's neighbors in the adjacency graph. Selecting a node from the mini-map brings up the inspection view for the indicated node.

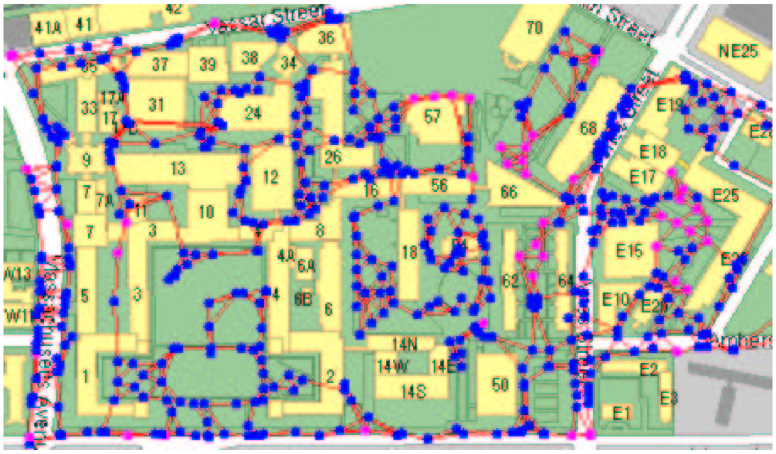
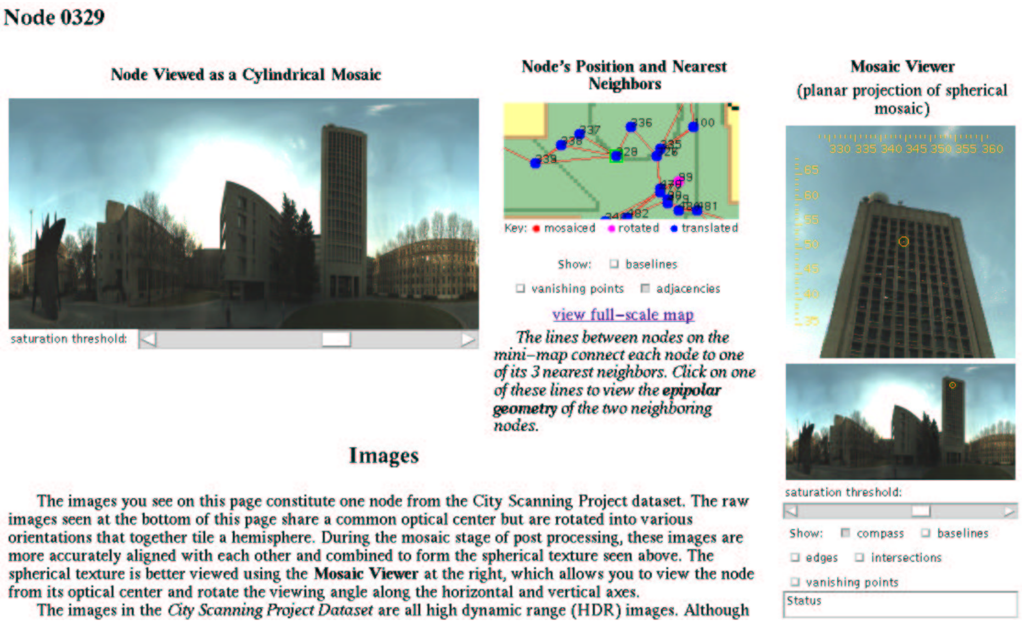


Figure 7: The web interface to geo-referenced nodes (points), adjacencies (edges) and map.

Selecting an edge from the mini-map brings up an epipolar geometry view for the implicated node pair (Figure 9). This view depicts each node position as a small cross in the counterpart node. The user can indicate a point in either node, and see the point's epipolar line (ruled with metric tick marks) displayed in the adjacent node.



Node 0329

Node Viewed as a Cylindrical Mosaic

Node's Position and Nearest Neighbors

Mosaic Viewer

(planar projection of spherical mosaic)

The lines between nodes on the mini–map connect each node to one of its 3 nearest neighbors. Click on one of these lines to view the ***epipolar geometry*** *of the two neighboring nodes.*

Images

The images you see on this page constitute one node from the City Scanning Project dataset. The raw images seen at the bottom of this page share a common optical center but are rotated into various orientations that together tile a hemisphere. During the mosaic stage of post processing, these images are more accurately aligned with each other and combined to form the spherical texture seen above. The spherical texture is better viewed using the **Mosaic Viewer** at the right, which allows you to view the node from its optical center and rotate the viewing angle along the horizontal and vertical axes.

The images in the *City Scanning Project Dataset* are all high dynamic range (HDR) images. Although

Figure 8: The web interface to one node, with a cylindrical mosaic (left), an interactive perspective view of the spherical mosaic (right), and a "mini-map" of the node's context (center).

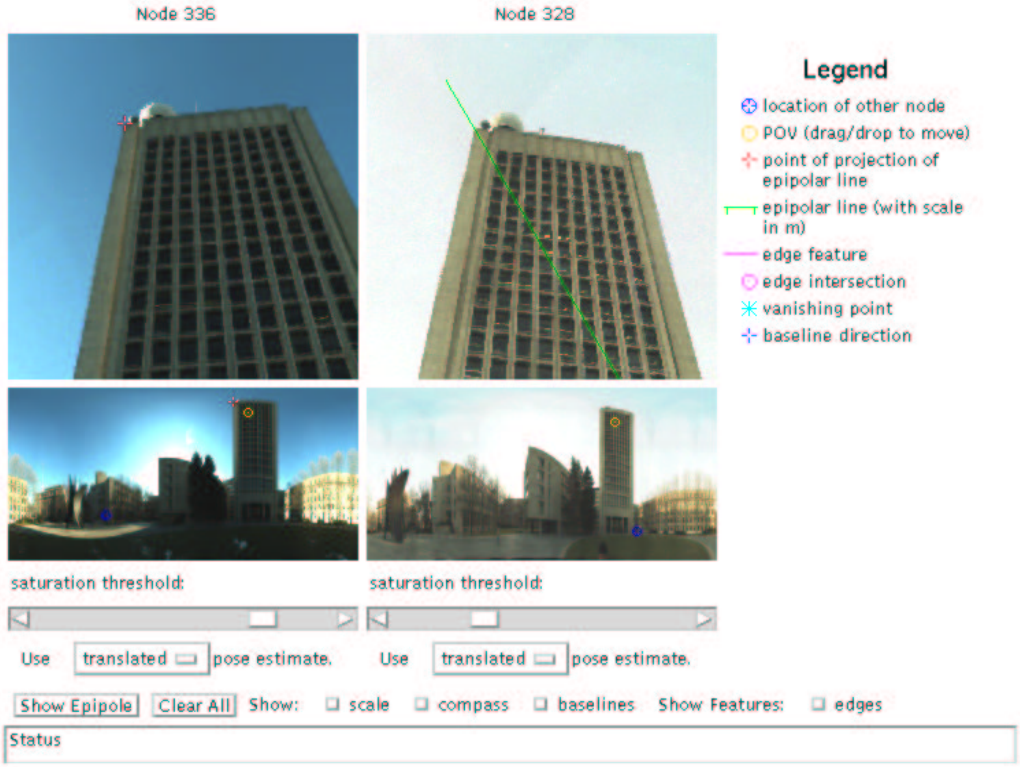


Figure 9: The web interface to the epipolar geometry of a pair of nodes.

5 Existing Datasets and Related Work

This section briefly reviews existing systems and methods for acquiring geo-referenced image datasets.

5.1 Satellite-Based (Robotic) Acquisition

A number of robotic mapping systems (e.g., [NAS]) incorporate satellites which continuously acquire high-altitude imagery of Earth. These systems provide a wealth of data about regions with limited vertical relief (oceans, much natural terrain, etc.). Since detailed information is maintained about each satellite's orbital parameters, the images acquired can be geo-referenced fairly accurately by the sensor itself; one data interchange site for publically available imagery [Geo] states registration to within about twelve meters on the surface of the Earth. Commercial systems, and presumably classified systems, provide even higher accuracy.

These systems are less useful for imaging high-relief regions such as cities. Here, a satellite at a great distance from the scene can acquire only near-nadir views (in which case near-vertical surfaces are imaged very obliquely) or near-horizon views (in which case most near-vertical surfaces are largely occluded). For urban environments, a near-ground sensor is necessary to acquire unoccluded, nearly fronto-parallel views of these surface. Autonomous low-altitude flying vehicles (e.g. [SDF$^+$98]) exist, but have not yet been demonstrated to acquire accurately geo-referenced imagery.

5.2 Interactive (Semi-Automated) Methods

Another route to acquiring near-ground, geo-referenced imagery is through manual interaction. A variety of semi-automated methods have been proposed for recovering exterior parameters for small image sets, in applications for photogrammetry [Wol74, Sla80, Gre97], digital mapping, and computer graphics scene modeling [BB95, DTM96, SHS98]. These systems combine automated or semi-automated feature detection with semi-automated bundle adjustment, in which a human operator indicates or selects corresponding point features across multiple images. Sometimes, geo-referenced points (e.g., painted crosses) are placed in the scene before the sensor is deployed. In this case, the human user can geo-reference the imagery by associating features visible in the image to known features in an existing geo-referenced feature set.

Semi-automated methods are fundamentally limited in a number of respects. First, these methods are scale-limited; the number of person-hours required to process a dataset of more than a few hundred images would be prohibitive in most situations. Human operators typically rely on every pair of images overlapping in some fashion, so that common elements can be indicated; in extended datasets, most image pairs are completely unrelated due to occlusion. Second, interactive methods are vulnerable to human failings: errors and short-cuts. Humans may make errors by indicating incorrect matches in ambiguous situations (for example in the presence of occlusion and visual clutter). Also, we have observed that human operators tend to specify only as many constraints as are required for nominal convergence of the underlying optimization. These practices generate insufficient or erroneous constraints, and unstable bundle adjustment solutions. Finally, we note that semi-automated methods do not scale with underlying technology (i.e., CPU speed), but rather have the human operator, whose throughput is essentially fixed, as their bottleneck. Thus, the throughput of a semi-automated system will improve little over time.

5.3 Summary

The state of existing work can be summarized as follows. Although there are other sensor platforms that produce geo-referenced imagery, they produce data that is not suitable for close-range urban mapping. Similarly, although interactive bundle adjustment techniques for image registration exist, they do not scale well to the huge image datasets needed for modeling extended areas. Prior to the acquisition system used to collect the datasets described in this paper, no scalable, automated system had been demonstrated to acquire close-range, accurately geo-referenced imagery of urban areas.

6 Conclusion

We described an intrinsically and extrinsically calibrated terrestrial image dataset acquired within an extended region on the MIT campus. The dataset is available on-line in a format that supports interactive browsing and download.

The acquisition system that produced this dataset operates under the assumption that its sensor has been deployed in an urban area, one exhibiting at least two vanishing points in every omnidirectional observation, and point features that are persistently visible under camera motions of a few meters. The dataset contains many images, acquired over a large area. The exterior calibration information associated with the images is self-consistent to a few centimeters of position and a fraction of a degree of orientation. Due to its scale and extent, achieving comparable datasets using current semi-automated methods would require substantial manual effort. Our automated system, in contrast, enables accurate, large-scale image registration.

We have made this data available to the research community in the hope that it will be useful to researchers and developers of large-scale image self-calibration and registration algorithms, image-based rendering and metric 3D reconstruction from calibrated image datasets, and digital cartography and GIS applications.

A Data Conventions and Formats

This appendix describes image, calibration (intrinsic and extrinsic), and feature data at each stage of system processing. We also describe the format of a number of data files used within the system to represent these elements, and the organization of these files in the on-line repository.

A.1 Coordinate Information

A coordinate information file resides at the top of the data hierarchy. It specifies an external coordinate system for reference, and the dataset's origin and coordinate axes expressed in this coordinate system. Here, for example, is the coordinate information file for one of our datasets:

```
CITY_LOCAL_TANGENT_PLANE
DATUM WGS84
LTP_LATITUDE_DEG 42.363371136
LTP_LONGITUDE_DEG -71.090968114
LTP_ALTITUDE_M 46.41
LTP_TO_ECEF_XFORM_ROW1 -0.67383016 -0.535169569 0.509457014 3255071.19
LTP_TO_ECEF_XFORM_ROW2 0.738886267 -0.488049937 0.464601273 2968474.63
LTP_TO_ECEF_XFORM_ROW3           0  0.689493141 0.724292212 4596736.62
LTP_TO_ECEF_XFORM_ROW4 0 0 0 1
```

The DATUM element specifies the WGS84 ECEF (Earth-centered, Earth-fixed) datum [HWLC97] as an external reference. The first three LTP fields specify the latitude, longitude, and altitude of our coordinate system origin (in this case, a GPS base station on the roof of our building) with respect to the base datum, in degrees and meters respectively. Finally, the four fields LTP_TO_ECEF_XFORM_ROWi specify the rows of a 4×4 matrix that converts a column vector $(x, y, z, 1)^T$, expressed in LTP coordinates, to ECEF coordinates.

In ECEF coordinates (which are right-handed), the origin is at the center of the reference ellipsoid; the positive x axis pierces the equator and the prime meridian; the positive z axis pierces the North pole; and the positive y axis is orthogonal to the x and z axes. In LTP coordinates (also right-handed), the origin is specified in ECEF coordinates, and $\hat{z}$ points away from the center of the WGS84 ellipsoid (in Cambridge, this direction deviates roughly 1.7 milliradians from the gravity vector). Finally, LTP $\hat{x}$ and $\hat{y}$ are the projections of due East and due North directions, respectively, into the "local level plane" normal to $\hat{z}$.

A.2 Image Descriptor Files

The image data is uploaded to the laboratory servers in a directory named according to the date and time of the start of the acquisition run, and in a file named according to the date, time and approximate position of the image shuttering. All images are stored in a lossless RGB format. With each acquired image, the sensor associates an image descriptor file and a pose descriptor file. At the time of shuttering, an image descriptor file is produced by the sensor process that controls the pan-tilt head and shutters the camera. This file captures all shuttering-specific information (except the camera's intrinsic and extrinsic parameters), including:

- A header field CITY_INFO;
- Digital camera identifier (make and model);
- Date and time (GMT) of image acquisition;
- The source of the image (sensor or program);
- The image type (radiance) and pixel type (log-radiance);

- Image width and height in pixels, and number of color channels (3);
- Exposure bracketing and photometric calibration information; and
- The lens focal length and aperture setting used.

Here is an example image descriptor file produced by our sensor:

```
CITY_INFO
GMT_YEAR        2000
GMT_MONTH       4
GMT_DAY         25
GMT_HOUR        20
GMT_MINUTE      2
GMT_SECOND      14
IMAGE_TYPE      radiance-image
IMAGE_WIDTH     1300
IMAGE_HEIGHT    1030
DEPTH           3
PIXEL_TYPE      log-radiance-map
NUM_AVERAGED    3
NUM_EXPOSURES   5
EXPOSURE_TIMES  9.70E-5 1.27E-2 2.53E-2 5.06E-2 1.01E-1
MAX_RADIANCE    3.6183085441589355
MIN_RADIANCE   -0.7229903340339661
CAMERA_TYPE     Wintriss-1300ASC
APERTURE        16.0
LENS            8.5mm
GAMMA1         -0.1875479966402054
GAMMA2          0.2141740024089813
CAMERA_SPECIFIC
```

The image files store only eight bits per color channel. We convert the color channel value, stored in the image file, into a radiance value proportional to the flux of light coming into the camera as:

$$r \quad \propto \quad \exp\left[\left(\left(p/255 * \left(\mathtt{RAD}_{\mathtt{MAX}} - \mathtt{RAD}_{\mathtt{MIN}}\right)\right) + \mathtt{RAD}_{\mathtt{MIN}}\right) - 2\ln\left(f\right)\right],$$

effectively scaling radiance by the reciprocal of the aperture squared.

Raw values of 0 and 255 are used as sentinels to mark undersaturated and saturated pixels, respectively. There are few such pixels in the data; they occur only when the camera observes very dark areas, or very bright specular reflections, or the sun itself.

A.3 Camera Pose Descriptor Files

For each image, a camera pose descriptor file is logged by a separate process that controls the sensor navigation system. This file captures what is known of the camera's intrinsic and extrinsic parameters at the time of each image acquisition. This file includes fields describing:

- Digital camera identifier (make and model);
- The source of the image (sensor or program);
- Image width, height (in pixels);
- Camera focal length (in pixels);
- Camera principal point c_x, c_y;
- Camera skew (assumed zero);
- Camera LTP position (x, y, z);
- Camera orientation (q_0, q_1, q_2, q_3).

Here is an example pose descriptor file produced by our sensor:

```
CITY_CAMERA     Wintriss-1300ASC
SOURCE          ARGUS
WIDTH           1299
HEIGHT          1027
FOCAL_X         1192.14
FOCAL_Y         1197.48
SKEW            0
CENTER_X        627.627
CENTER_Y        487.751
TRANSLATION     265.491 -371.936 -42.213
ROTATION        0.0418 0.064 -0.696 0.714
```

The translation field represents the position of the camera's optical center, expressed in LTP coordinates. We represent the rotation field as a quaternion that, when converted to matrix form [Sho85], expresses the rotation that takes coordinates expressed in world (LTP) coordinates into the camera coordinate system. In camera coordinates, the Z axis is aligned along the positive optical axis of the camera (i.e., Z increases into the image), X goes from left to right when looking through the camera, and Y increases downward on the image.

Each subsequent processing stage that modifies the rotation or translation fields appends tokens to the file describing the result of processing. For example, upon successful termination, the `mosaic` process writes two lines to each pose descriptor file:

```
MOSAIC_STATUS   CONVERGENT
MOSAIC_RESIDUE  0.029940
```

where the status field reports convergence, and the residue field is as described in Section 3.1.

A.4 Node Descriptor Files

At each sensor position, a collection of images is acquired about a common optical center. We call this image collection a “node.” Each image is indexed, zero-relative, with respect to the “base image,” i.e. the first image acquired in each node. Each node has a “node descriptor file,” containing the number of images in the node, and the index of the base image. Here is a node descriptor file produced by the sensor:

```
CITY_NODE
NUM_IMAGES 20
BASE_IMAGE 0
```

With each node is also associated a “mosaic adjacency graph” file listing, for each image index, those images that have significant overlap with this one in the mosaic tiling. Here is an example adjacency file for a 20-image tiling (comment lines begin with #):

```
# Automatically generated adjacency graph
# Number after central image, ordered by overlap area
0 :   1  7 15
1 :   0  2 14
2 :   1  3 13
[additional node adjacency descriptors omitted]
```

A.5 Feature Descriptor Files

The system uses edge features and intersection-based point features, both localized to sub-pixel precision and stored in ASCII format. Here is an example edge feature descriptor file:

```
XRES YRES
NEDGES
BEGIN_EDGE 56
POINTS          x1 276.035248 y1 133.566696 x2 197.981735 y2 136.868851
COV             c11 0.00 c12 0.00 c13 0.00 c22 0.00 c23 0.00 c33 0.00
LEN             78.123329
MAG             15.086859
LINE            a 0.042269 b 0.999106 c 145.114929
END_EDGE
[additional edge feature descriptors omitted]
```

Here is an example point feature descriptor file:

```
XRES YRES
NPOINTS
BEGIN INTERSECTION
EDGES 2
ID                  650
IMAGE_X             112.5921173096
```

```
IMAGE_Y             361.2429504395
EDGE_MAGNITUDE      42741.8007812500
LENGTH              43024.5351562500
ANGLE              -1.5103152990
...
END INTERSECTION
[additional point feature descriptors omitted]
```

A.6 Vanishing Point Descriptor Files

The rotational registration stage detects the scene vanishing points (VP) observed by each node. The results are stored in a VP descriptor file. The VP direction is expressed in node coordinates, and the VP variance (i.e., width at 95% confidence) is expressed in degrees squared. `NumPoints` refers to the number of image features combined to produce the VP direction estimate. Here is an example VP descriptor file for a single node.

```
Num 4
VP -0.994017225652078 -0.041513024196319 0.101026847565489
 Variance 0.000172442672632687
 NumPoints 54
[additional node VP descriptors omitted]
```

After rotational registration of all nodes to a common coordinate system, a vanishing point descriptor file is produced for the entire dataset. This file describes the union of all vanishing points observed by all nodes in the dataset. Here is an example global VP descriptor file, with VP directions expressed in LTP coordinates, and variances in degrees squared:

```
Num 8
VP 0.880051924246545 0.474578089570595 0.0168596420328816
 Variance 2.61946183099412E-05
[additional global VP descriptors omitted]
```

Acknowledgements

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (contract DACA76-97-K-0002), and of the Office of Naval Research (MURI Award SA 1524-2582386).

References

[AT00] Matthew Antone and Seth Teller. Automatic recovery of relative camera rotations for urban scenes. In *Proc. CVPR*, pages II–282–289, June 2000.

[AT01] Matthew Antone and Seth Teller. Scalable, absolute position recovery for omni-directional image networks. In *Proc. CVPR*, pages I–398–405, December 2001.

[AT02] Matthew Antone and Seth Teller. Scalable extrinsic calibration of omni-directional image networks. *IJCV*, 49(2/3):143–174, Sept./Oct. 2002.

[BB95] Shawn Becker and V. Michael Bove. Semiautomatic 3-D model extraction from uncalibrated 2-D camera views. In *Proc. Visual Data Exploration and Analysis II, SPIE Vol. 2410*, pages 447–461, 1995.

[BdT99] Michael Bosse, Douglas de Couto, and Seth Teller. Eyes of Argus: Georeferenced imagery in urban environments. *GPS World*, pages 20–30, April 1999.

[Bin74] Christopher Bingham. An antipodally symmetric distribution on the sphere. *Annals of Statistics*, 2(6):1201–1225, Nov. 1974.

[CMT98] Satyan Coorg, Neel Master, and Seth Teller. Acquisition of a large pose-mosaic dataset. In *CVPR '98*, pages 872–878, 1998.

[CT99] Satyan Coorg and Seth Teller. Extracting textured vertical facades from controlled close-range imagery. In *Proc. CVPR '99*, pages 625–632, June 1999.

[CT00] Satyan Coorg and Seth Teller. Spherical mosaics with quaternions and dense correlation. *IJCV*, 37(3):259–273, 2000.

[DeC98] Douglas DeCouto. Instrumentation for rapidly acquiring pose imagery. Master's thesis, Dept. of Electrical Engineering and Computer Science, MIT, 1998.

[DM97] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '97 Conference Proceedings*, August 1997.

[DTM96] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH '96 Conference Proceedings*, pages 11–20, August 1996.

[Geo] GeoTIFF, `http://www.remotesensing.org/geotiff/geotiff.html`.

[Gre97] C.W. Greeve. *Digital Photogrammetry: an Addendum to the Manual of Photogrammetry.* American Society of Photogrammetry and Remote Sensing, 1997.

[HWLC97] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *GPS Theory and Practice.* Springer-Wien, 1997.

[Lar98] G.W. Larson. LogLuv encoding for full-gamut, high-dynamic range images. *Journal of Graphics Tools*, 3:15–31, 1998.

[NAS] NASA's Earth Observing System, `http://eospso.gsfc.nasa.gov/earth_observ.html`.

[NSD] National Spatial Data Infrastructure, `http://www.fgdc.gov/nsdi/nsdi.html`.

[Ope] Open GIS Consortium, `http://www.opengis.org`.

[PS85] Franco P. Preparata and Michael Ian Shamos. *Computational Geometry: an Introduction.* Springer-Verlag, 1985.

[Ros84] A. Rosenfeld, editor. *Multiresolution image processing and analysis.* Springer-Verlag, 1984.

[SDF$^+$98] C.P. Sanders, P.A. DeBitetto, E. Feron, H.F. Vuong, and N. Leveson. Hierarchical control of small autonomous helicopters. In *Proc. 37th IEEE Conf. on Decision and Control*, Dec. 1998.

[Sho85] Ken Shoemake. Animating rotation with quaternion curves. *Computer Graphics (Proc. Siggraph)*, 19(3):245–254, 1985.

[SHS98] H. Shum, M. Han, and R. Szeliski. Interactive construction of 3-d models from panoramic mosaics. In *Proc. CVPR*, pages 427–433, 1998.

[Sla80] C.C. Slama. *Manual of Photogrammetry.* American Society of Photogrammetry and Remote Sensing, 1980.

[TAB$^+$01] Seth Teller, Matthew Antone, Michael Bosse, Satyan Coorg, Manish Jethwa, and Neel Master. Calibrated, registered images of an extended urban area. In *Proc. CVPR*, December 2001.

[Tel97] Seth Teller. Automatic acquisition of hierarchical, textured 3D geometric models of urban environments: Project plan. In *Proc. Image Understanding Workshop*, pages 767–770, May 1997.

[Tel98] Seth Teller. Automated urban model acquisition: Project rationale and status. In *Proc. Image Understanding Workshop*, pages 455–462, Nov. 1998.

[Ter] TerraServer, the Vertical Portal for Overhead Imagery, `http://www.terraserver.com`.

[Wol74] P.R. Wolf. *Elements of Photogrammetry.* McGraw-Hill, 1974.

[WTT$^+$02] Xiaoguang Wang, Stefano Totaro, Franck Taillandier, Allen Hanson, and Seth Teller. Recovering facade texture and microstructure from real-world images. In *Proc. 2nd International Workshop on Texture Analysis and Synthesis at ECCV*, pages 145–149, June 2002.

[Zha98] Z. Zhang. A flexible new technique for camera calibration. Technical Report MSR-TR-98-71, Microsoft Research, Dec. 1998.