Scalable, Controlled Imagery Capture in Urban Environments

Seth Teller
MIT Laboratory for Computer Science

Abstract

We describe the design considerations underlying a system for scalable, automated capture of precisely controlled imagery in urban scenes. The system operates for architectural scenes in which, from every camera position, some two vanishing points are visible. It has been used to capture thousands of controlled images in outdoor environments spanning hundreds of meters. The proposed system architecture forms the foundation for a future, fully robotic outdoor mapping capability for urban areas, analogous to existing, satellite-based robotic mapping systems which acquire images and models of natural terrain.

Four key ideas distinguish our approach from other methods. First, our sensor acquires *georeferencing metadata* with every image, enabling related images to be efficiently identified and registered. Second, the sensor acquires *omni-directional images*; we show strong experimental evidence that such images are fundamentally more powerful observations than conventional (narrow-FOV) images. Third, the system uses a *probabilistic, projective error formulation* to account for uncertainty. By treating measurement error in an appropriate depth-free framework, and by deferring decisions about camera calibration and scene structure until many noisy observations can be fused, the system achieves superior robustness and accuracy. Fourth, the system's computational requirements scale *linearly* in the input size, the area of the acquisition region, and the size of the output model. This is in contrast to most previous methods, which either assume constant-size inputs or exhibit quadratic running time (or worse) asymptotically.

These attributes enable the system to operate in a regime of scale and physical extent which is unachievable by any other method, whether manual or automated. Consequently, it can acquire the most complex calibrated terrestrial image sets in existence, while operating faster than any existing manual or algorithmic method.

1. Introduction

For some years, robotic mapping systems have been in orbit above Earth [1]. These systems include satellites which continuously acquire digital images and other data, and report these data along with information about the instantaneous position of the acquiring satellite to a number of ground stations which fuse the observations into useful representations or models of the Earth's surface. A wide variety of mapping capabilities are provided by such sensors and systems, including the recovery of 3D models of natural terrain exhibiting limited vertical relief from multiple overlapping images [14]. However, satellite-based systems can not provide high-fidelity 3D models of urban regions, due to the high resolution images required, the extreme vertical relief of urban regions and the resulting inability of satellites to occupy useful vantage points, the absence of established fiducial or control points, and the wide variations in observed material properties which characterize urban regions in contrast to natural terrain.

In analogy to satellite-based mapping systems, our long-term goal is to develop a close-range robotic mapping capability for urban (i.e., man-made) environments, in which multiple, autonomous terrestrial sensors can be deployed throughout the region of interest to acquire a model of the region's 3D structure and appearance. As part of this effort, we have developed a collection of scaleable, automated model capture techniques that acquire and register near-ground imagery, and extract detailed 3D scene models from the registered images.

This paper describes the design principles underlying our system, which captures 3D geometric models of extended outdoor scenes directly from thousands of photographs. We designed the system to capture urban scenes extending over hundreds or thousands of meters, even in the presence of real-world occlusion, visual clutter, and lighting variations. Our system includes a semi-robotic sensor and a fully automated set of algorithms for metric camera registration and 3D reconstruction. In its design, we have also laid the groundwork for a future, fully robotic controlled image acquisition and model capture capability (for example by multiple, cooperating autonomous ground and air sensors). This paper gives a high-level view of the system and the key ideas behind it; more detailed descriptions of its components can be found elsewhere [37, 38, 11, 12, 13, 2, 7, 3, 39].

Our model capture system includes a number of innovations, such as a geo-referenced camera, high-

resolution omni-directional imagery, and a suite of robust, asympotically efficient algorithms for image registration in the presence of uncertainty (i.e., sensor and feature detection noise). These innovations together enable the system to handle large-scale problems. Moreover, because it is automated, the system's throughput grows with that of its underlying sensor and computers, unlike semi-automated systems which, having a human operator as their bottleneck, exhibit essentially constant performance over time.

For *small* datasets of a few dozen images and a few buildings, interactive modeling tools (e.g., [16]) are presently the most effective technique today for end-to-end model construction. In this operating regime, the human operator can infer and (in concert with an interactive user interface) indicate 3D shape, correspondences for bundle adjustment, and surface texture more effectively and accurately than can any computer vision algorithm at present.

The conventional view has been that this state of affairs holds true at every problem scale, so that scaling to large instances, even if achievable, can not be achieved without an intolerable sacrifice of accuracy. We show that in fact this tradeoff *does not hold* for model capture. In particular, we demonstrate that as the scale of the modeling task grows, automated systems gain a decisive advantage, achieving greater speed *and* greater accuracy than a human operator for essential tasks such as camera registration. For example, our system readily registers, in a few CPU-hours and to greater accuracy, image datasets which would require tens or hundreds of hours of manual effort with any semi-automated registration method, and which cannot be handled by any previously proposed algorithmic method.

2. Key Ideas

Four key ideas distinguish our approach from previous methods. Two involve augmenting a conventional camera to provide fundamentally more powerful sensor observations (Sections 2.1 and 2.2). The other two ideas concern the system's algorithmic treatment of geometric uncertainty and scale (Sections 2.3 and 2.4).

2.1. Imagery and Pose Metadata

Our sensor acquires approximate *geo-referencing metadata* with every image, that is, an estimate of the camera's "pose," or 6-DOF position and orientation, in an absolute coordinate system. This pose metadata, even though only approximate, is critical to achieving both scaling and end-to-end automated operation.

2.1.1. Implications for Scaling

Intuitively, pose metadata enables scaling by making it possible for the system to identify images which are likely to have observed common or overlapping portions of the scene, without examining every pair of images for possible correlation. Position estimates accurate to within a few meters suffice for this purpose; this accuracy is achievable today with inexpensive GPS receivers.

Pose metadata enables both parallel image acquisition by multiple cameras, and asympotically efficient determination of overlapping image sets, regardless of the number of cameras. To see this, imagine n cameras positioned throughout a scene, each simultaneously acquiring an image. Without pose metadata, any algorithm processing the n images would be forced to inspect all pairs of images to find common elements for calibration, registration, or 3D reconstruction, and would therefore expend at least $O(n^2)$ time. Moreover, most of this effort would be wasted, since occlusion would cause most pairs of images to have no pixels in common.

Perhaps surprisingly, an analogous argument applies to the case of a *single* camera acquiring n images in an uninterrupted sequence, even under the strong assumption that all sub-sequences of constant length observe overlapping scene elements. For in any real system, measurements are noisy, and derived quantities such as camera pose or reconstructed feature positions accumulate error over time. Thus, any arbitrarily long image sequence will require comparisons between non-adjacent images (i.e., images separated by more than a constant offset) to achieve registration of the entire sequence.

In other words, without absolute pose, each new image must be compared to every previous image to detect path crossings, resulting in quadratic processing complexity. This is why algorithms which track long video sequences (e.g. [21]) require that a human operator inform the system of whether the sequence is "open" or "closed," and if the latter, supply the offset of that image with which the algorithm should attempt to correlate the first acquired image. Note too, that the very assumption that images have been acquired in a single sequence amounts to a scaling limitation, since it excludes the possibility of parallel image capture by multiple cameras operating simultaneously, or one camera deployed in several distinct sessions.

In contrast, when pose information is available for each image, the system can efficiently determine those images acquired in or near any region, simply by inserting all images into a spatial index as they are acquired, then performing a proximity or inverse-range query for the region [32] to discover all nearby images. These images, by virtue of their mutual proximity to the query region, are likely to contain overlapping observations.

Moreover, when images are acquired with roughly constant spatial density, we can expect that the number of images adjacent to a given image will be constant, and that the number of images acquired inside a query region will be a function only of the region's size (i.e., area or volume). The total size of the adjacency graph for the images will therefore be linear in n, and under the constant density assumption the time required to construct the graph will also be linear in n.

2.1.2. End-to-End Automation

Pose metadata enables automation and end-to-end operation (i.e., the production of controlled imagery directly from sensor data) in two ways, both of which involve the elimination of a task which is traditionally performed interactively by a human operator.

First, pose metadata, even if only approximate, can serve as initial values during photogrammetric bundle adjustment, removing the need for manual initialization by a human operator, for example by specifying approximate camera positions in preparation for subsequent numerical optimization ([9, 16]).

The claims above and in Section 2.1.1 hold for any *a priori* pose metadata, even if it is supplied only in a scene-relative coordinate system (e.g., if all cameras report their pose with respect to some fiducial coordinate system observable in the scene). However, when image pose metadata is expressed in Earth-relative coordinates [25], as in our system, a further advantage becomes apparent. In this case, the controlled imagery can itself be expressed in Earth-relative coordinates, making it readily integrable into existing image, terrain, and GIS (geographical information systems) datasets. This eliminates the need for a human operator to supply or indicate photogrammetric "tie points" as required by conventional systems [35, 24] to register newly controlled imagery to existing Earth-relative datasets.

2.2. Omni-Directional Imagery

Our sensor acquires wide-FOV *omni-directional images*[1] [31, 36], rather than conventional narrow-FOV (planar) images. Omni-directional imagery provides a number of fundamental advantages, theoretically, numerically, and operationally.

2.2.1. Theoretical Advantages

Narrow-FOV images exhibit the aperture problem [41], due to which small camera rotations and small camera translations are indistinguishable to first order in some circumstances. Moreover, egomotion and estimation algorithms for conventional imagery can exhibit bias, by searching for and reporting motion solutions that lie in or slightly outside the camera's field of view, regardless of their true locations.

In contrast, omni-directional imagery is free of the aperture problem and its attendant biases. A number of researchers have investigated motion solutions on the sphere, reasoning that a wider field of view enables more robust estimation of global motion patterns. For example, because omni-directional imagery captures observations in antipodal pairs, small rotations and small translations can be robustly distinguished on the sphere ([23, 19, 38]). Similarly, because omni-directional imagery imposes no preferred direction on the image observations, motion estimation is free of directional bias.

2.2.2. Practical Advantages

These are more than theoretical considerations. In practice, spherical images form fundamentally more powerful observations than conventional images, for a variety of reasons. First, spherical imagery makes self-calibration of conventional imagery more robust and accurate; when several conventional images are known to share the same optical center, panoramic mosaics can be constructed more reliably and with less error, and intrinsic parameters can be estimated more accurately, by enforcing the constraint that all images are related by pure rotations, and that images form an adjacency graph which tiles the sphere [13].

Second, since a spherical image observes more of the scene than a conventional image, when scene quantities are aggregated across the entire field of view, they mutually enforce the statistical peaks indicative of scene quantities (for example, in a Hough Transform [27]) more strongly than do conventional images. Thus scene structures such as vanishing points can be detected reliably [2] where related techniques, invoked on conventional images, fail [10]. This is true even accounting for decreased resolution, that is, when the same number of pixels are used for conventional and spherical imagery.

Third, for noisy feature observations, the additional field of view afforded by spherical imagery enables the aggregation of a greater number of observations, and thus more accurate estimation [2].

Finally, the ability to search a larger field of view enables feature matching algorithms to succeed even under extremely challenging conditions (up to 80% outliers, or as low as 20% feature overlap, in our experiments on synthetic data, and with baselines of tens of meters for real data acquired outdoors [3].

2.2.3. Operational Advantages

Spherical images are advantageous from an operational standpoint as well. In any system using conventional, narrow-FOV imagery, the camera operator or sensor

[1] Also called "full-view panoramas" or "spherical images" even though in practice they view only part of a sphere.

platform must keep a particular subject structure in view during image acquisition (for example, by manually and continuously reorienting the camera [21]) to ensure scene tracking and continuous camera egomotion recovery. This requirement amounts to another kind of scaling limitation: since severe occlusion or clutter may make it impossible to keep a given subject structure in view, the system can not be deployed in general scenes, or in situations in which the identity of the subject structure is unknown at the time of acquisition.

Thus spherical imagery provides an important practical advantage in the future realization of fully robotic (i.e., autonomous) model capture systems, by removing a restriction on the sensor operator or platform. This reduces the complexity of the computations that must occur on-board the sensor (or in the case of our current system, the complexity of the decisions that must be made by the sensor's human operator). In this sense, the use of spherical imagery largely decouples the deployment of the sensor from the processing of the observations acquired by the sensor. (This decoupling is not complete, since clearly in order to acquire observations of some region of interest, the sensor must be dispatched to the vicinity of the region.)

2.3. Projective, Probabilistic Models

Our system accounts for *geometric uncertainty* at every system stage using a projective, probabilistic ("soft") error formulation. This has a number of advantages, including appropriate modeling and fusion of noisy measurements, and the deferral of "hard" decisions (for example, about the existence of scene structure) until many individual observations can be combined. This probabilistic formation also allows the system to handle an unknown number of match features, unknown occlusion, deocclusion, and outliers during image registration.

2.3.1. Projective Uncertainty Models

Image observations are formed by projection onto an imaging surface, and therefore contain no *a priori* depth or metric distance information. We use a projective uncertainty model to represent and fuse observations in the absence of depth information. For example, the uncertainty of line and point features is modeled on the sphere of directions, rather than in 3D. Since the rotation which registers one node to another is represented as a unit quaternion, its uncertainty is modeled on the four-sphere. For both tasks, we use Bingham distributions [6], essentially Gaussian distributions conditioned to lie on the sphere.

2.3.2. Projective Data Fusion

Image observations, and the geometric features derived from them, are inherently noisy. Many system components combine, or fuse, many noisy measurements in order to estimate a smaller number of quantities more accurately. In the case of image registration (i.e., extrinsic camera calibration), this "projective data fusion" enables estimation of aggregate features with an error far lower than that contained in any single feature observation. For example, projective data fusion of hundreds of thousands of noisy low-level features produces extremely accurate estimates of high-level "ensemble features" such as vanishing points and the focus of motion expansion. One contribution of the system is a principled, implemented projective data fusion method for line features [2] and point features derived from line intersections [3].

Projective, "soft" data fusion enables the system to defer decisions about camera registration and scene structure until many local and semi-local observations have been fused. In this way it achieves superior robustness and accuracy over techniques restricted to local observations (for example, tracking algorithms based on registering successive frames).

2.4. Linear Asymptotics

In our system, all algorithms run in *linear time* in the number of input images, the size of the acquisition region, and the complexity of the output. This is in contrast to previous methods, which either assume constant-size inputs or expend quadratic time (or worse). Linear asymptotics enables the system to operate in a regime of sheer scale and physical extent which is unachievable by any other approach at present.

Efficient algorithms also facilitate overdetermined formulation of derived scene quantities such as feature and camera positions, enabling them to be recovered more accurately and stably than can underconstrained quantities. Finally, linear running times allow the system to employ the heuristic strategy of liberal low-level feature detection. In other words, we tolerate spurious low-level features; these tend not to produce spurious high-level conclusions, since they are not reinforced by multiple independent observations. This has the practical advantage that the numerical parameters unavoidably associated with feature detection can be set liberally, removing the need for most parameter tuning in the system.

3. System Overview

This section gives a high-level view of the system architecture and implementation. We also describe system development and data acquisition strategies of interest. Finally, we briefly describe our long-term goals for the

evolution of the system.

3.1. Processing Stages

The controlled image acquisition system comprises two processing stages: sensing and registration. Sensing (photography, image acquisition) is the acquisition of image observations. Registration (image exterior control, extrinsic calibration, bundle adjustment) is the process of bringing images into pixel-accurate alignment. There are many applications of accurately controlled imagery, including model extraction and image-based rendering. Here we focus on sensing and registration only.

The sensing phase largely involves the deployment of sensor hardware, a digital camera augmented with navigation instrumentation. The registration stage involves the execution of a series of software algorithms, organized as a batch (automated) processing pipeline.

3.2. End-to-End Operation

There are many algorithmic techniques which address one part of the machine vision problem, for example camera self-calibration, egomotion recovery, feature matching, geometry and texture reconstruction, etc. In practice, every implementation comes with an associated set of limitations; for example, egomotion recovery may work only for short image sequences; feature matching may work only for a small number of features, or under controlled or nearly uniform illumination, or only over short baselines; etc. Algorithms often rely on numerical parameters which are understood only empirically, so that from a newly encountered dataset it is difficult or impossible to determine operational parameter values.

All of these factors, in addition to the asymptotic limitations previously discussed, make it difficult to compose existing techniques to achieve an end-to-end capability. For example, one can not in general successfully deploy an egomotion recovery technique outdoors if it has been designed to operate under controlled or constant illumination. Similarly, one can not apply feature matching algorithms to long-baseline sequences, when the algorithms were designed to assume short baselines.

In contrast, we have devoted particular effort to architecting a system which is end-to-end in that it transforms, in a completely automated series of software stages, raw input images into controlled output images in geodetic coordinates. Our application focuses on architectural scenes; implying that the system must process image data acquired outdoors over long baselines, with significant occlusion, under uncontrolled illumination (i.e., daylight), and in the presence of severe visual clutter due to extraneous scene elements. In our setting, the usual assumption of visual overlap among frames holds, but only locally; as the sensor is moved far from any reference position, the fraction of the scene commonly in view drops to zero.

Successful function in our current setting, under these operating conditions, is critical to the eventual realization of a practical robotic mapping system capable of deployment over extended outdoor areas under general illumination and in general settings. We note too that, in order to be truly end-to-end, the system must produce geo-referenced data to enable its automated incorporation into existing GIS-based data systems.

At present, our system makes an assumption about scene structure: that in every (omni-directional) image, two or more distinct architectural vanishing points, or families of parallel lines, are visible. Our current image registration algorithms rely on these vanishing points. We observe that many egomotion recovery algorithms assume persistent point features in the scene; we assume them as well, with an additional requirement that two or more features be located at infinity (and therefore that they be insensitive to translations). The vanishing point assumption holds for more than 95% of the images acquired in and around our urban campus [39].

3.3. Breadth-First Development

We followed a breadth-first strategy of data collection and system architecture and development. In contrast to depth-first development technique in which a series of individual modules are elaborated in turn, we elaborated the system's information flow architecture, and a prototype module for each component in parallel, before elaborating any single system module in depth. Also, we commenced data collection at the start of the project, using an early sensor prototype. During data collection we made no attempt to select unoccluded views or particularly favorable lighting conditions (other than to avoid darkness and inclement weather). As a result, our datasets include highly cluttered and occluded images acquired under varying illumination, and immediately present challenges of scale and generality.

This data collection and development strategy forced us to address issues of scale and input generality from the start, and ensured the elaboration of modules on an as-needed basis as dictated by the challenges of calibrating, registering, and extracting structure from noisy, complex real-world image and pose data.

3.4. Toward Robotic Mapping

Our long-term goal is to achieve a fully robotic image acquisition and mapping capability, in which one or more cooperating autonomous sensors can be deployed, on the ground or in the air, in the vicinity of a region

to be modeled. Each of the key ideas in the system's development are motivated by this long-term goal.

Omni-directional imaging reduces the real-time computational load on each sensor, and enables more robust egomotion recovery for the images collected by each sensor. Pose metadata enables efficient fusion of image data from any number of sensors. Our probabilistic, projective error model enables the appropriate treatment of the noisy real-world metadata, images and derived features reported by the sensors. Linear asymptotics is critical to realizing a computationally practical model capture capability at very large scales. And of course, full automation is essential by definition if we are to achieve a truly autonomous, robotic acquisition capability.

3.5. Future Directions

Most camera egomotion recovery algorithms assume persistent (i.e., trackable) scene structure, as does our approach. We make the additional assumption of at least two point features lying at infinity (vanishing points). Many, but not all, urban scenes exhibit vanishing points; this assumption enables us to decouple the 6-DOF image registration problem into two lower-dimensional problems, each of which can be solved efficiently. We are now removing the vanishing point assumption, and exploring whether robust, large-scale egomotion recovery is still possible, for example through factored optical flow on the sphere.

The system currently requires a human operator to move the sensor about the subject environment. We anticipate that in the future the human operator will be supplanted by semi-autonomous or autonomous acquisition vehicles (e.g. [29, 33]). As these sensors come on-line, the annotated imagery they acquire can be input directly into our system for processing, to realize a robotic, terrestrial controlled image capture capability.

4. Related Work

This section describes a variety of other proposed approaches to acquiring terrestrial calibrated imagery.

4.1. Algorithmic Methods

A wide variety of algorithms have been proposed in photogrammetry [40] and computer vision [26, 18]. Algorithmic methods typically address a sub-problem of the end-to-end model extraction problem, such as camera calibration or registration, feature detection or matching, structure extraction, or texture estimation or inverse global illumination. Due to the difficulty of the general problem, algorithm designers formulate their methods under some set of assumptions about the attributes, quality, or scale of the input to be provided. One measure of a technique's utility is its paucity of assumptions, i.e., the generality of the setting in which the algorithm is applicable.

Some algorithms assume special camera configurations in order to ease the work of matching or tracking points across frames. Examples include closely-spaced stereo pairs or triples [17, 34], linear gantries [8], or (in small-scale laboratory settings) circular camera paths arranged by placing the subject of acquisition on a turntable [20, 30]. Other methods assume a smooth image sequence (e.g, from a slowly moving video camera), with the camera manually reoriented so as to keep a subject building in view [4, 21], so that consecutive (temporally adjacent) frames can be assumed to view highly overlapping scene geometry. Some methods assume "closed" video sequences, for which the camera is moved in a circuit to arrive at or near its original position and orientation [21], effectively informing the system that the first and last frames in the sequence view common scene geometry. Extraction methods may assume significant standoff from the scene to ensure that buildings fall within the camera's limited field of view, or assume (or select) unoccluded and/or uncluttered views. Other methods observe only a few sides of the scene, resulting in a partial, facade-like model of the observed structures.

Each of these assumptions implies a limitation on the applicability of the method to the general model extraction problem. Algorithms may be defeated by widely spaced cameras, cameras which move so that scene objects repeatedly come in and out of view, "open" image sequences (or closed sequences not known to be so), or significant changes in lighting or specular effects. Stability problems can arise when attempting to propagate local information throughout a global framework.

Combinatorial limitations may be in effect as well. Methods that search for common features (points, edges, texture patches) in all pairs of images expend time quadratic in the number of images. Combining model geometry produced by methods that operate in private coordinate systems (for example, on sets of images acquired by different cameras in the same scene) may require combinatorially prohibitive searching, or a human operator to indicate corresponding scene elements. The assumed organization of images has implications for system throughput, as well: any algorithm that treats its input as a single image sequence related only through image adjacency must process every image in the sequence (with at least linear time delay) before it can relate the first image to the last image.

In the absence of a priori navigation information, scene fiducials of known dimension, or "tie points" with known geodetic coordinates, scene recovery algorithms operate in an arbitrary 3D coordinate system, recover-

ing scene structure only up to an arbitrary scale, orientation and translation.

Operating assumptions also make it difficult to compose algorithms into working end-to-end systems, since the output of one algorithm may not be useful, or suitable, as the input to another. Finally, when algorithms are composed it becomes more difficult to acquire datasets which meet the increased restrictions that arise. For example, it may not be possible to acquire an image dataset that has both large extent (i.e., covers a large spatial area) and is acquired under constant illumination conditions, simply because outdoor illumination conditions change over time.

4.2. Interactive Methods

In view of the computational cost and fidelity of these approaches, researchers have developed interactive programs which rely on a human operator to perform tasks which ease the burden of the computer vision algorithms in the system [5, 16, 22]. Interactive methods tend to be more comprehensive, or end-to-end. Broadly speaking, these methods decouple the problem's combinatorial and grouping aspects, which are delegated to the human operator, from its optimization aspects, which are performed algorithmically [16].

A good exemplar of hybrid modeling tools is Facade [16], which requires a human to perform image acquisition, site preparation (establishing a coordinate system), coarse layout of building block structure, camera pose initialization, indication of low-level and high-level features and feature correspondences, and segmentation of images into subject and clutter. (Manual segmentation may be difficult or impossible in the presence of complex clutter such as foliage.)

Although interactive methods can produce high-quality models given enough human effort, these methods exhibit several fundamental limitations. First, the system cannot function without a skilled operator. Second, the scale and extent of datasets which can be manipulated interactively is fundamentally limited by the human operator's capacity and skill level (excluding, for example, extended urban areas or visually cluttered scenes). Third, interactive systems are "non-algorithmic" in the sense that they can not be subjected to standard performance measures such as asymptotic time and space requirements. Fourth, the human operator will eventually be the bottleneck of the system, as the speed of underlying processor technology increases, but the human capacity does not. Finally, it is difficult to parallelize interactive systems, since considerable communication would be necessary among different operators both to partition the input and merge the resulting outputs.

5. Common Questions

Other researchers have raised two classes of questions about the system and design principles described here.

5.1. Aren't Interactive Tools Sufficient?

One frequently raised question is, "Good interactive tools exist to solve the model capture problem from images. Why pursue an automated approach?"

Indeed useful tools exist. However, there are two reasons why these tools do not solve the model capture problem.

First, these tools require an enormous amount of manual effort, even for simple models. Modeling a single bell tower to moderate detail, and twenty surrounding buildings as simple block models, required about 80 man-hours (two man-weeks) of effort by a skilled user of Facade [15]. Modeling a significant portion of the Los Angeles Basin, about 1,500 city blocks with 15,000 buildings, required about 100,000 man-hours (*fifty man-years*) [28] in an effort by UCLA's urban visualization project using less sophisticated tools. Interestingly, in both efforts, the amount of manual time required was roughly 4-6 man-hours per building, regardless of the modeling software used! Clearly it is not feasible to expend this level of effort routinely.

Second, a model produced from a small number of images is *fundamentally different* from a model produced from many images, in that it is "view-biased": it supports high-quality synthetic rendering only from a limited set of viewpoints that are near those of the source cameras, or observe portions of the model that the source cameras observed at high resolution. This is acceptable for applications in which the set of viewpoints to be exercised is known in advance; in particular for animation sequences. However, it is not acceptable for applications requiring a freely (e.g., interactively) controlled viewpoint, such as in virtual reality or architectural exploration or visualization.

How does a small number of images bias the resulting viewpoint? The most obvious example is resolution: the captured model can be imaged at high resolution only where a source camera acquired an image at that resolution. Second is self-occlusion; if a model self-occludes, than with only a small number of source images, much of the model will never be observed, and will therefore have to be hallucinated during view synthesis. Third is directionality of reflection characteristics. Suppose the real-world building has significant specular surfaces. Unless each such surface is observed from sufficiently many directions to recover its directionally-dependent reflection characteristics with some confidence, synthetic views from directions other than those employed by the source cameras will not be faithful to the original.

These observations lead to the following conclusion: to view a captured object faithfully, in synthesis, from a variety of distances and directions, it must be photographed sufficiently closely, and from sufficiently many directions, to allow high-confidence reconstruction. We estimate that roughly 100,000 images are required per square kilometer of building area to capture the region to square-centimeter resolution ($1\ \text{km}^2 = 10^{10}\ \text{cm}^2 = 10{,}000$ images with a mega-pixel camera, with another factor of 10 for stereo and pixel redundancy). This number of images is far beyond the capacity of one (or many cooperating) human users to process in a reasonable time.

5.2. Why Not Use a Range-Finder?

Another frequently raised question is, "Why not use an active laser range finder, rather than a passive camera, to gather data? This will produce depth estimates directly, rather than indirectly through correlation, correspondence, etc." We have formulated several responses to this question, each on different grounds.

First, the extraction of geometric information from images is a hard, interesting, and long-standing problem in photogrammetry and machine vision. This alone is sufficient reason to pursue it.

Second, there are often operating and engineering constraints which preclude the use of active sensors in a particular application. The most obvious of these is the requirement in some military scenarios for stealth. An active sensor will by design broadcast its location, which is unacceptable in some settings. In practice, range finders may not operate well in bright sunlight, and may be slow, heavy, power-hungry, etc. The bottom line is that the choice of *any* sensor, camera or ranger, will dictate or constrain an enormous number of engineering considerations; it simply doesn't make sense to say that one sensor is better than the other except with respect to a particular set of operating requirements.

Finally, we argue that the choice of sensor is largely *independent of* the algorithmic ideas embedded in our particular systems approach. The four key ideas described earlier (position metadata; wide field of view; linear asymptotics; and probabilistic feature processing) are useful *regardless of the type of sensor used*. Imagine deploying a range sensor over many city blocks; each of the problems tackled in our context (scaling, matching, registration) would arise in this context as well, and can be addressed using the ideas outlined above.

6. Contributions and Conclusions

We described a number of design principles applicable to techniques for automated capture of controlled terrestrial imagery.

First, we propose the use of geo-referenced cameras which annotate each acquired image with approximate position and orientation metadata. We show that this enables asympotically faster, and fundamentally more robust, image registration algorithms, and removes the need for human initialization of bundle-adjustments. Absolute pose estimates also enable parallel image acquisition, and absolute (geo-referenced) output, enabling the output models to be merged automatically with existing GIS data.

Second, we discuss experimental evidence that omni-directional images are fundamentally more powerful observations than are conventional (narrow-FOV) images. These images enable much more robust and accurate camera calibration and registration than existing methods. Moreover, they decrease the processing that must occur on future autonomous sensor platforms during acquisition.

Third, we model uncertainty at every stage of the system, from low-level feature detection to camera registration to scene reconstruction. We show that by treating uncertainty appropriately, in a projective (depth- and distance-free) framework, observational noise can be overcome at large scales to produce accurate camera registration. The resulting system robustly handles datasets with many images and features over large acquisition regions.

Fourth, we demonstrate a suite of algorithms, with asymptotically linear time and space usage in the size of the input and output. These methods enable efficient operation at large scaling regimes, and allow the system to achieve constraint sets which are highly overdetermined, overcoming significant real-world lighting variation, visual clutter, and ambiguity.

Finally, and consequently, we show that the conventional tradeoff – increase automation, decrease accuracy – does not hold for the image control task in our capture system.

Acknowledgements

Support for this research was provided in part by the Department of Defense Advanced Research Projects Agency under contract DACA76-97-K-0002, the Office of Naval Research under MURI Award #1524-2582386, and by Intel Corporation.

References

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

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

[3] M. Antone and S. Teller. Scalable, absolute position recovery for omni-directional image networks. In *Proc. CVPR (to appear)*, 2001.

[4] P. A. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure from motion. *Lecture Notes in Computer Science*, 800:85–96, 1994.

[5] S. Becker and V. M. 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.

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

[7] M. Bosse, D. de Couto, and S. Teller. Eyes of argus: Georeferenced imagery in urban environments. *GPS World*, pages 20–30, April 1999.

[8] R. Collins. A space-sweep approach to true multi-image matching. In *Proc. CVPR*, pages 358–363, 1996.

[9] R. Collins, C. Jaynes, F. Stolle, X. Wang, Y. Cheng, A. Hanson, and E. Riseman. A system for automated site model acquisition. In *Proc. SPIE Vol. 7617*, 1995.

[10] R. T. Collins and R. Weiss. Vanishing point calculation as statistical inference on the unit sphere. In *Proc. ICCV*, pages 400–403, Dec. 1990.

[11] S. Coorg, N. Master, and S. Teller. Acquisition of a large pose-mosaic dataset. In *CVPR '98*, pages 872–878, 1998.

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

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

[14] R. Crippen and R. Blom. Imageodesy: A tool for mapping subpixel terrain displacements in satellite imagery. Technical report, International Union of Geodesy and Geophysics, Geophysics and the Environment, July 1995.

[15] P. Debevec. Personal communication, Feb. 2001.

[16] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *SIGGRAPH '96 Conference Proceedings*, pages 11–20, Aug. 1996.

[17] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *ECCV92*, pages 563–578, 1992.

[18] O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.

[19] C. Fermüller and Y. Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *Int'l Journal of Computer Vision*, 28(2):137–154, 1998.

[20] A. Fitzgibbon, G. Cross, and A. Zisserman. Automatic 3D model construction for turn-table sequences. *Lecture Notes in Computer Science*, 1506:155–170, 1998.

[21] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. *Lecture Notes in Computer Science*, 1406:311–326, 1998.

[22] D. P. Gibson, N. W. Campbell, and B. T. Thomas. The generation of 3-D models of outdoor objects from uncalibrated still images. In *Image Processing and its Applications*, pages 28–32. Institute of Electrical Engineers, London, July 1999.

[23] J. Gluckman and S. Nayar. Ego-motion and omnidirectional cameras. In *ICCV*, pages 35–42, 1998.

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

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

[26] B. K. P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[27] P. V. C. Hough. A method and means for recognizing complex patterns. U. S. Patent No. 3,069,654, 1962.

[28] W. Jepson. Personal communication, Feb 2000.

[29] D. Kang, J. Anderson, and P. DeBitetto. Draper unmanned vehicle systems. In *Proc.* 3^{rd} *Intelligent Autonomous Vehicles*, March 1998.

[30] P. R. S. Mendonça and R. Cipolla. Estimation of epipolar geometry from apparent contours: Affine and circular motion cases. In *Proc. CVPR*, pages 9–14. IEEE, Jun 1999.

[31] S. Nayar. Catadioptric omnidirectional camera. In *Proc. CVPR*, pages 482–488, 1997.

[32] F. P. Preparata and M. I. Shamos. *Computational Geometry: an Introduction*. Springer-Verlag, 1985.

[33] C. Sanders, P. DeBitetto, E. Feron, H. Vuong, and N. Leveson. Hierarchical control of small autonomous helicopters. In *Proc.* 37^{th} *IEEE Conference on Decision and Control*, Dec. 1998.

[34] A. Shashua. Trilinearity in visual recognition by alignment. In *ECCV94*, pages A:479–484, 1994.

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

[36] R. Szeliski and H. Shum. Creating full-view panoramic mosaics and texture-mapped 3D models. In *SIGGRAPH '97 Conference Proceedings*, pages 251–258, Aug. 1997.

[37] S. Teller. Automatic acquisition of hierarchical, textured 3D geometric models of urban environments: Project plan. In *Proc. the Image Understanding Workshop*, 1997.

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

[39] S. Teller, M. Antone, M. Bosse, S. Coorg, M. Jethwa, and N. Master. Calibrated, registered images of an extended urban area. In *Proc. CVPR (to appear)*, 2001.

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

[41] G. S. Young and R. Chellappa. Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field. *IEEE PAMI*, 14(10):995–1013, Oct. 1992.