Towards Urban 3D Reconstruction From Video

A. Akbarzadeh*, J.-M. Frahm+, P. Mordohai+, B. Clipp+, C. Engels*, D. Gallup+,
P. Merrell+, M. Phelps*, S. Sinha+, B. Talton+, L. Wang*, Q. Yang*, H. Stewenius*,
R. Yang*, G. Welch+, H. Towles+, D. Nistér* and M. Pollefeys+

*Department of Computer Science
Center for Visualization and Virtual Environments
University of Kentucky, Lexington, USA

+Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, USA

Abstract

The paper introduces a data collection system and a processing pipeline for automatic geo-registered 3D reconstruction of urban scenes from video. The system collects multiple video streams, as well as GPS and INS measurements in order to place the reconstructed models in geo-registered coordinates. Besides high quality in terms of both geometry and appearance, we aim at real-time performance. Even though our processing pipeline is currently far from being real-time, we select techniques and we design processing modules that can achieve fast performance on multiple CPUs and GPUs aiming at real-time performance in the near future. We present the main considerations in designing the system and the steps of the processing pipeline. We show results on real video sequences captured by our system.

1 Introduction

Detailed, 3D models of cities are usually made from aerial data, in the form of range or passive images combined with other modalities, such as measurements from a Global Positioning System (GPS). While these models may be useful for navigation, they provide little additional information compared to maps in terms of visualization. Buildings and other landmarks cannot be easily recognized since the façades are poorly reconstructed from aerial images due to bad viewing angles. To achieve high-quality ground-level visualization one needs to capture data from the ground. A system that automatically generates texture-mapped, ground-level 3D models should be capable of capturing large amounts of data while driving through the streets and of processing these data efficiently.

In this paper, we introduce an approach for fully automatic 3D reconstruction of urban scenes from several hours of video data captured by a multi-camera system. The goal is an automatic system for processing very large amounts of video data acquired in an unconstrained manner. This forces us to take shape from video out of the laboratory and to achieve a fieldable system.

The video acquisition system consists of eight cameras mounted on a vehicle, with a quadruple of cameras looking to each side. The cameras have a resolution of 1024×768 pixels and a frame rate of 30 Hz. Each quadruple consists of cameras directed straight sideways (orthogonal to the driving direction), and diagonally forward, backward and upwards with minimal overlap to achieve a large horizontal and vertical field of view. Additionally, the acquisition system employs an Inertial Navigation System (INS) and a GPS to enable geo-registration of the cameras. Examples of ground-level reconstructions from our system can be seen in Figs. 1 and 2.

Figure 1. Example of dense reconstruction.

The entire acquisition system is packaged in a sealed pod, which is mounted on the back of a vehicle. As the vehicle is driven through urban environments, the captured

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

Figure 2. Dense reconstruction of a city block.

video is stored on disk drives in the pod. After a capture session, the drives are moved from the pod on the vehicle to a 10-PC (dual-processor) computer cluster for processing. Our performance goal is to process up to 6 hours of acquired data in an equal amount of time.

Processing entails the following steps: sparse reconstruction during which the geo-registered poses of the cameras are estimated from the video and the INS/GPS data; and dense reconstruction during which a texture-mapped, 3D model of the urban scene is computed from the video data and the results of the sparse step.

In sparse reconstruction the trajectory of the camera is estimated from the video data using structure from motion techniques. The goal is to achieve precise camera poses in order to support temporal multi-view stereo, while keeping a globally coherent geo-registered trajectory free of drift. To this end, the INS/GPS data are post-processed to obtain a filtered precise trajectory of the vehicle, which is called Smoothed Best Estimated Trajectory (SBET). The SBET and the hand-eye calibration between the origin of the SBET coordinate system and the coordinate systems of the cameras provide reliable estimates of the camera trajectories.

In dense reconstruction, the surfaces of the buildings, ground and other structures are estimated using multi-view stereo techniques. The goal of this step is to provide accurate surfaces wherever possible even in the presence of ambiguous or little surface texture, occlusion or specularity. The reconstruction step is divided into multi-view stereo, which produces depth-maps from multiple views with a single reference view, and depth-map fusion, which resolves conflicts between multiple depth maps and derives a coherent surface description. The dense reconstruction stage also provides texture for the surfaces using the video input.

The remainder of the paper is organized as follows. Section 1.1 discusses related work. The processing pipeline is described in detail in Section 2, while the different system aspects of a multi-camera capture system with INS/GPS recording are outlined in Section 3. Experimental results are reviewed in Section 4 with conclusions in Section 5.

1.1 Previous Work

The research community has devoted a lot of effort to the modeling of man-made environments using a combination of sensors and modalities. Here, we briefly review work relying on ground-based imaging since it is more closely related to our project. An equal, if not larger, volume of work exists for aerial imaging. The typical goal is the accurate reconstruction of urban or archaeological sites, including both geometry and texture, in order to obtain models useful for visualization, quantitative analysis in the form of measurements at large or small scales and potentially for studying their evolution through time.

A natural choice to satisfy the requirement of modeling the geometry and appearance is the combined use of active range scanners and digital cameras. Stamos and Allen [1] used such a combination, while also addressing the problems of registering the two modalities, segmenting the data and fitting planes to the point cloud. El-Hakim et al. [2] propose a methodology for selecting the most appropriate modality among range scanners, ground and aerial images and CAD models. Früh and Zakhor [3] developed a system that is very similar to ours since it is also mounted on a vehicle and captures large amounts of data in continuous mode, in contrast to the previous approaches that captured a few, isolated images of the scene. Their system consists of two laser scanners, one for map construction and registration and one for geometry reconstruction, and a digital camera, for texture acquisition. A system with similar configuration, but smaller size, that also operates in continuous mode was presented by Biber et al. [4]. Other work on large scale urban modeling includes the 4D Atlanta project carried out by Schindler et al. [5], which also examines the evolution of the model through time. Cornelis et al. [6] have also developed a system specialized for the reconstruction of façades from a stereo rig mounted on a moving vehicle.

Laser scanners have the advantage of providing accurate 3D measurements directly. On the other hand, they can be cumbersome and expensive. Several researchers in photogrammetry and computer vision address the problem of reconstruction relying solely on passive sensors (cameras) in order to increase the flexibility of the system while de-

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

creasing its size, weight and cost. The challenges are due mostly to the well-document inaccuracies in 3D reconstruction from 2D measurements. To obtain useful models one may have to interact with the system or make simplifying assumptions. Among the first such attempts was the MIT City Scanning project, an overview of which can be found in [7]. A semi-automatic approach under which simple geometric primitives are fitted to the data was proposed by Debevec et al. [8]. Compelling models can be reconstructed even though fine details are not modeled but treated as texture instead. Rother and Carlsson [9] show that multiple-view reconstruction can be formulated as a linear estimation problem given a known fixed plane that is visible in all images. This approach also requires manual operations. Dick et al. [10] presented an automatic approach that infers piecewise planar surfaces from sparse features taking into account constraints such as orthogonality and verticality. The authors later proposed a more elaborate, MCMC-based method [11] that uses generative models for buildings. It is also fully automatic, but is restricted by the prior models and can only operate on small sets of images, typically two to six. Similar high-level reasoning is also employed by [5]. Werner and Zisserman [12] presented an automatic method, inspired by [8], that fits planes and polyhedra on sparse reconstructed primitives by examining the support they receive via a modified version of the space sweep algorithm [13].

We approach the problem using passive sensors only, building upon the experience from intensive study of structure from motion and shape reconstruction within the computer vision community in the last two decades. Since this literature is too large to survey here, the interested reader is referred to [14, 15]. The emphasis in our project is on developing a fully automatic system that is able to operate in continuous mode without the luxury of capturing data from selected viewpoints since capturing is performed from a moving vehicle constrained to the vantage points of urban streets. Our system design is also driven by the performance goal of being able to post-process the large video datasets in a time equal to the acquisition time. Our assembled team has significant experience in most if not all aspects of structure from motion and stereo processing involved in producing textured, 3D models from images and video [16, 17, 18, 19, 20, 21].

2 Processing Pipeline

In the following we describe the different techniques used in our system in more detail. The processing pipeline begins by estimating a geo-registered camera pose for each frame of the videos. We approach this by determining 2D-2D point correspondences in consecutive video frames. Then, we use the relative camera geometry of the internally calibrated cameras to establish a Euclidean space for the cameras. The INS/GPS information is used to compute the camera position in the geo-spatial coordinate system.

Once the camera poses have been computed, we use them together with the video frames to perform stereo matching on the input images. This leads to a depth map for each frame. These depth maps are later fused to enforce consistency between them. A flow chart of the processing pipeline is shown in Fig. 3.

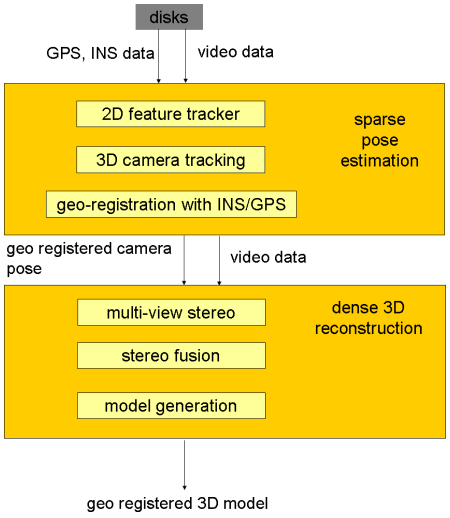


Figure 3. 3D processing pipeline

2.1 2D Feature Tracking

To establish 2D feature correspondences between consecutive video frames we track features with a hierarchical KLT tracker [22]. To achieve real-time tracking with video frame rate we use an implementation of the hierarchical KLT tracker on the GPU [23]. It needs on average 30ms to track 1000 feature points in a 1024×768 image on an ATI X1900 graphics card.

The weakness of tracking techniques are large disparity ranges as the flow-assumption of motion of less than a pixel at the corresponding pyramid level limits the amount of motion that can be captured. Thus video frames with large disparities pose problems to the KLT tracker. Hence, we can also use a detect and match tracker similar to [24]. Its strength is that it can search large disparity ranges very quickly – faster than video can be fetched from disk. Its weakness is that in noisy, low-texture conditions the repeatability of detection is not always reliable (a phenomenon similarly noted by [25]).

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE
IEEE COMPUTER SOCIETY

2.2 3D Camera Tracking

We are investigating and developing several approaches to determine the camera pose from the 2D feature tracks, depending on the availability of INS/GPS data. We would like our system to be functional in the absence of such data.

When INS/GPS data are not available, we use a vision-only camera tracking algorithm along the lines of [18]. Briefly stated, we can initialize the camera tracker with the relative pose of three views, given feature correspondences in them. These correspondences are triangulated using the computed camera poses. Additional poses are computed with RANSAC and hypothesis-generation using constraints from 2D feature to 3D world point correspondences. New world points are re-triangulated using new views as they become available.

To avoid accumulated drift the system is periodically re-initialized with a new set of three views. We stitch the new poses into the old coordinate system exploiting the constraints of one overlapping camera. The remaining degree of freedom is the scale of the old and the new coordinate system. It is estimated using corresponding triangulated points in both coordinate frames.

All pose estimation methods use preemptive RANSAC and local iterative refinement for robustness [26]. In practice, the system must re-initialize frequently unless we use bundle adjustment to refine poses. With bundle adjustment the pose estimation is less sensitive to measurement noise which leads to fewer re-initializations.

2.3 Geo-Registration with INS/GPS Data

To determine geo-registered coordinates of the features in the 3D model, we employ the INS/GPS data. The INS/GPS measurement system is outfitted with a GPS receiver, gyroscopes, and accelerometers. It delivers highly accurate measurements of the position and orientation of the vehicle on which the cameras are mounted.

A Euclidean transformation, which will be referred to as the hand-eye calibration, maps the center of the geo-location system to the optical center of each of the cameras. Initially each camera keeps its own coordinate frame. The optical center of the first frame of each camera is the origin and the optical axis and the axes of the first image plane are used as the axes. The scale is arbitrarily chosen by setting the distance between the first and second camera positions in the video sequence to unit length.

Our first implementation of geo-registration computes a similarity transformation (rotation, translation and scale) between the poses of a given camera in the vision coordinate system and the poses of each camera in the world. This approach has difficulties in dealing with drift in the vision-based camera pose estimation since it is limited to one rigid transformation for all frames.

We are currently developing a second approach which overcomes these limitations by fusing geo-location measurements and tracked 2D features either using a Kalman filter or through bundle adjustment. These methods are expected to outperform the similarity transformation, geo-registration technique.

2.4 Multi-View Stereo

The multi-view stereo module takes as input the camera poses and images from a single video stream and produces a depth map for each frame. It uses the plane-sweep algorithm of Collins [13], which is an efficient multi-image matching technique. Conceptually, a plane is swept through space in steps along a predefined direction, typically parallel to the optical axis. At each position of the plane, all views are projected on it. If a point on the plane is at the correct depth, then, according to the brightness constancy assumption, all pixels that project on it should have consistent intensities. We measure this consistency by summing the absolute intensity differences in square aggregation windows defined in the reference image, for which the depth map is computed. The hypothesis with the minimum cost (sum of absolute differences) is selected as the depth estimate for each pixel.

(a) (b)

Figure 4. Stereo depth maps rendered as 3D models. (a) Fronto-parallel sweep only. (b) Multiple sweeping directions.

Traditional plane-sweeping techniques typically sweep fronto-parallel planes only, which do not account for perspective observed in non-fronto-parallel surfaces. We extend the algorithm by sweeping planes in multiple directions, where the directions are aligned with the planar surfaces we expect to observe in the scene such as the ground and building façades. We can deduce the orientations of the ground and façade planes beforehand by assuming the vehicle drives parallel to the ground and to the façades, and that the façades are vertical and meet at right angles. Figure 4 illustrates the improvements gained by sweeping in multiple directions. Besides its ability to process multiple images at the same time, the plane-sweep stereo algorithm can easily

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

be ported to the GPU to achieve very fast performance [27].

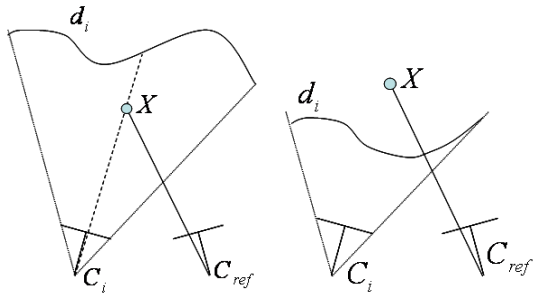
2.5 Stereo Fusion

Multi-view stereo provides a depth map for every reference frame. Since we do not enforce consistency between depth maps during stereo, we need to enforce it in a separate stereo fusion step. Fusion serves two purposes: it improves the quality of the depth estimates by ensuring that estimates for the same point are consistent with multiple depth maps and it produces more economical representations by merging multiple depth estimates into one. Related work includes volumetric [28, 29], patch-based [30], viewpoint-based [31] and pixel-based [32] approaches. We opt for a viewpoint-based approach inspired by the median fusion algorithm [17]. A volumetric method is impractical since it would require a very large number of voxels for our image resolution and accuracy requirements. Instead we operate on the set of depth hypotheses for each pixel of the reference view. It is useful to have the image plane as a reference, because then the final mesh can be generated rapidly by triangulating the depth estimates of adjacent pixels.

Given depth maps from a set of consecutive frames, the stereo fusion step resolves conflicts between computed depths and produces a depth map in which most of the noise has been removed. Fusion is highly effective because points are visible in large numbers of frames, which provide multiple depth estimates for each point. Even though each depth estimate is produced by a simple, fast stereo algorithm, the consensus among these estimates is usually very accurate. Fusion is designed to resolve conflicts like those illustrated in Figure 5. In Figure 5.a, the depth estimate X in the view of the reference camera C_{ref} occludes the depth map d_i of camera C_i, while in Figure 5.b, the depth estimate X of the reference camera passes the depth map of camera P_i. Both situations are impossible and should be corrected.

One approach is the median fusion algorithm of [17]. The input is a set of depth maps, one of which, typically the one in the middle, is used as reference. The algorithm aims at selecting the best depth estimate for each pixel based on their *stability*, which depends on the number and type of conflicts between each depth hypothesis and other depth maps. For each depth estimate of the reference depth map, the number of other depths maps that it occludes (Figure 5.a) or passes (Figure 5.b) is computed. This process is repeated for each depth estimate of all other depth maps with respect to reference depth map. The selected (most stable) solution is defined as the minimum depth for which there are at least as many depth maps that occlude it as there are that pass it.

We are also working on a similar approach that takes into account the confidence of each depth estimate. Stereo produces more accurate depth estimates in parts of the image where there is more texture and no occlusion. The presence of unambiguous matches is indicated by a single strong minimum of the cost computed during the plane-sweep stage. On the other hand, the cost functions of ambiguous matches have multiple minima. Based upon the form of the cost function, we can assign a confidence value to each depth estimate. During fusion this confidence is increased or decreased depending on whether other depth estimates support or contradict it. Depth estimates with very low confidence are replaced with ones that agree more closely with their more confident neighbors. This approach introduces a bias towards smooth surfaces, which may be justified since our emphasis is on reconstructing buildings, and it is very effective in areas where the majority of depth estimates are wrong, such as uniform walls and streets. On the other hand, smoothness comes at the expense of fine details, and the computation of confidence requires additional processing cycles. Our future work in stereo fusion will focus on a faster implementation of our algorithms and better methods for resolving occlusions using multiple views.



(a) X occludes d_i (b) X passes d_i

Figure 5. Illustration of conflicts between depth maps that have to be resolved by fusion

From the fused stereo depth maps, we generate a triangulated, texture-mapped, 3D model of the scene in geo-registered coordinates.

3 Video Acquisition System

The on-vehicle video acquisition system consists of two main sub-systems - an 8-camera digital video recorder (DVR) and an Applanix INS/GPS (model POS LV) navigation system. The DVR streams the raw images to disk, and the Applanix system tracks position and orientation so the 3D models produced in post-processing can be created in a common geo-registered coordinate system.

The DVR is built with eight Point Grey Research (PGR) Flea color cameras, with one quadruple of cameras for each side of the vehicle as shown in Figure 6. Each camera has

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

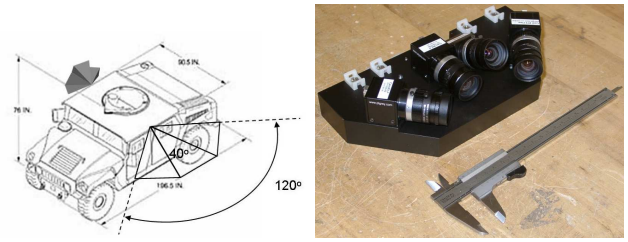


Figure 6. (a) Position of camera quadruples on the Humvee vehicle (b) Closeup of a camera quadruple.

a field-of-view of approximately $40° \times 30°$, and within a quadruple the cameras are arranged with minimal overlap in field-of-view. As shown in Figure 6, three cameras are mounted in the same plane to create a horizontal FOV of approximately 120 degrees. The fourth camera is tilted upward to create an overall vertical FOV of approximately 60 degrees with the side-looking camera.

The eight IEEE-1394 cameras are interfaced to eight Windows-Intel VME-based processor modules from Concurrent Technologies, each with a high-capacity SATA data drive. The eight-camera DVR is capable of streaming to disk 1024×768 Bayer-pattern images at 30Hz. The CCD exposure on all cameras is synchronized by using IEEE-1394 sync units from PGR.

With each video frame recorded, one of the DVR processor modules also records a GPS-timestamp message from the Applanix navigation system. These message events are also synchronized to the CCD exposure by means of an external TTL signal output by one of the cameras. In post-processing, this GPS timestamp is correlated with post-processed INS/GPS data to provide a smoothed best-estimate of the position and orientation of the navigation system and of each camera. The latter, of course, requires knowledge of the hand-eye calibration between the cameras and the INS/GPS system that one establishes during system calibration.

Camera Calibration Calibration of the cameras with respect to each other is desirable for fusing models created from independent streams of the eight cameras on the vehicle. Furthermore, calibration relative to the INS coordinate system is required in order to geo-register all the 3D models reconstructed from the captured video.

Camera intrinsics including lens distortion estimates are currently computed using a planar checkerboard and a toolkit built by Ilie and Welch [33] that is based on the OpenCV library. This solution provides a more automated calibration procedure that requires minimal user interaction.

Evaluation of the external relationship of cameras is complicated by the fact that there is little or no overlap in the visual FOV of the four cameras. The current method being pursued places the cameras within a large encompassing field of 3D feature points. The 3D feature points are actually created by an array of projectors, which illuminate a 3D display surface in front of the cameras with checkerboard patterns. The 3D position of each feature point is determined using a separately calibrated stereo camera pair.

Hand-eye (extrinsic) calibration of each camera quadruple relative to the INS must involve both camera quadruples and the navigation system mounted on the vehicle. Initially we plan to use calibrated feature points in the real-world whose position relative to each other has been established by theodolite. By viewing the feature points from numerous vehicle positions whose position and orientation are estimated by the Applanix system, we will then be able to establish the rotational-translational relationship of each camera with respect to the INS coordinate system. Bundle adjustment methods will undoubtedly be required. Future research will focus on removing the constraint of using pre-surveyed feature points.

4 Results

Figures 7 through 10 are illustrative of the fidelity of 3D models currently being reconstructed. All models were computed without any INS or GPS information except the one of Figure 10, which is geo-registered. The typical settings used for these reconstructions are the following: the number of features for tracking is 500, the number of images used for each depth map computation is 11, and the number of depth maps that are fused is 13.

Figure 7. Dense reconstruction from the forward camera.

IEEE COMPUTER SOCIETY

Figure 8. Dense reconstruction from the side camera.

Figure 9. Dense reconstruction from the side camera.

Figure 10. Geo-registered dense reconstruction from the side camera.

5 Summary and Conclusions

We have described a system aimed at real-time, dense, geo-registered, 3D urban reconstruction from video captured by a multi-camera system and INS/GPS measurements. The quality of the initial reconstructed results both with and without INS/GPS sensors is very promising. Future challenges include improving the accuracy of geo-registration, improving the process of bundle adjustment or Kalman filtering of the camera trajectory, registering and fusing the reconstructions across multiple video streams, speeding up the processing by porting operations to the GPU, and enhancing the processing pipeline to make it more robust. Potential longer term directions are change detection and the capability to perform incremental model updates using video acquired on different days.

Acknowledgement This work is partially supported by DARPA under the UrbanScape project, which is lead by the Geo-Spatial Technologies Information Division of SAIC. This document has been approved for public release, distribution unlimited, by DARPA.

References

[1] I. Stamos and P.K. Allen, "Geometry and texture recovery of scenes of large scale," *Computer Vision and Image Understanding*, vol. 88, no. 2, pp. 94–118, 2002.

[2] S.F. El-Hakim, J.-A. Beraldin, M. Picard, and A. Vettore, "Effective 3d modeling of heritage sites," in *4th International Conference of 3D Imaging and Modeling*, 2003, pp. 302–309.

[3] C. Früh and A. Zakhor, "An automated method for large-scale, ground-based city model acquisition," *Int. J. of Computer Vision*, vol. 60, no. 1, pp. 5–24, 2004.

[4] P. Biber, S. Fleck, D. Staneker, M. Wand, and W. Strasser, "First experiences with a mobile platform for flexible 3d model acquisition in indoor and outdoor environments – the waegele," in *ISPRS Working Group V/4: 3D-ARCH*, 2005.

[5] G. Schindler, P. Krishnamurthy, and F. Dellaert, "Line-based structure from motion for urban environments," in *3DPVT*, 2006.

[6] N. Cornelis, K. Cornelis, and L. Van Gool, "Fast compact city modeling for navigation pre-visualization," in *Int. Conf. on Computer Vision and Pattern Recognition*, 2006.

[7] S. Teller, "Automated urban model acquisition: Project rationale and status," in *Image Understanding Workshop*, 1998, pp. 455–462.

IEEE COMPUTER SOCIETY

[8] P. Debevec, C.J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," in *SIGGRAPH*, 1996, pp. 11–20.

[9] C. Rother and S. Carlsson, "Linear multi view reconstruction and camera recovery using a reference plane," *Int. J. of Computer Vision*, vol. 49, no. 2-3, pp. 117–141, 2002.

[10] A.R. Dick, P.H.S. Torr, and R. Cipolla, "Automatic 3d modelling of architecture," in *British Machine Vision Conference*, 2000, pp. 273–289.

[11] A.R. Dick, P.H.S. Torr, and R. Cipolla, "Modelling and interpretation of architecture from several images," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 111–134, 2004.

[12] T. Werner and A. Zisserman, "New techniques for automated architectural reconstruction from photographs," in *European Conf. on Computer Vision*, 2002, pp. 541–555.

[13] R.T. Collins, "A space-sweep approach to true multi-image matching," in *Int. Conf. on Computer Vision and Pattern Recognition*, 1996, pp. 358–363.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[15] O.D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.

[16] M. Pollefeys, R. Koch, and L. Van Gool, "Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters," *Int. J. of Computer Vision*, vol. 32, no. 1, pp. 7–25, 1999.

[17] D. Nistér, *Automatic dense reconstruction from uncalibrated video sequences*, PhD Thesis, Royal Institute of Technology KTH, Stockholm, Sweden, 2001.

[18] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–777, 2004.

[19] D. Nistér, "Automatic passive recovery of 3d from images and video," in *3DPVT*, 2004, pp. 438–445.

[20] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," *Int. J. of Computer Vision*, vol. 59, no. 3, pp. 207–232, 2004.

[21] M. Pollefeys, L. Van Gool, M. Vergauwen, K. Cornelis, F. Verbiest, and J. Tops, "Image-based 3d recording for archaeological field work," *Computer Graphics and Applications*, vol. 23, no. 3, pp. 20–27, 2003.

[22] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Int. Joint Conf. on Artificial Intelligence*, 1981.

[23] S. Sinha, J.-M. Frahm, and M. Pollefeys, "GPU-based Video Feature Tracking and Matching," Tech. Rep. TR06-012, University of North Carolina at Chapel Hill, May 2006.

[24] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, no. 1, 2006.

[25] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] D. Nistér, "Preemptive RANSAC for live structure and motion estimation," in *Int. Conf. on Computer Vision*, 2003, vol. 1, pp. 199–206.

[27] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," in *Int. Conf. on Computer Vision and Pattern Recognition*, 2003, pp. I: 211–217.

[28] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," *SIGGRAPH*, vol. 30, pp. 303–312, 1996.

[29] M.D. Wheeler, Y. Sato, and K. Ikeuchi, "Consensus surfaces for modeling 3d objects from multiple range images," in *Int. Conf. on Computer Vision*, 1998, pp. 917–924.

[30] P.V. Fua, "From multiple stereo views to multiple 3-d surfaces," *Int. J. of Computer Vision*, vol. 24, no. 1, pp. 19–35, 1997.

[31] P.J. Narayanan, P.W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," in *Int. Conf. on Computer Vision*, 1998, pp. 3–10.

[32] R. Koch, M. Pollefeys, and L. Van Gool, "Multi viewpoint stereo from uncalibrated video sequences," in *European Conf. on Computer Vision*, 1998, pp. 55–71.

[33] A. Ilie and G. Welch, "Ensuring Color Consistency across Multiple Cameras," in *Int. Conf. on Computer Vision*, 2005, pp. 1268–1275.

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY