

GENERATING THE SPARSE POINT CLOUD OF A CIVIL INFRASTRUCTURE SCENE USING A SINGLE VIDEO CAMERA UNDER PRACTICAL CONSTRAINTS

Fei Dai
Abbas Rashidi

Georgia Institute of Technology
130 Hinman Research Building
723 Cherry Street NW
Atlanta, GA 30332, USA

Ioannis Brilakis

Georgia Institute of Technology
328 Sustainable Education Building
788 Atlantic Drive NW
Atlanta, GA 30332, USA

Patricio Vela

Georgia Institute of Technology
TSRB 441 / Van Leer 368
Mail Code 0250
Atlanta, GA 30332, USA

ABSTRACT

Automating the model generation process of infrastructure can substantially reduce the modeling time and cost. This paper presents a method to generate a sparse point cloud of an infrastructure scene using a single video camera under practical constraints. It is the first step towards establishing an automatic framework for object-oriented as-built modeling. Motion blur and key frame selection criteria are considered. Structure from motion and bundle adjustment are explored. The method is demonstrated in a case study where the scene of a reinforced concrete bridge is videotaped, reconstructed, and metrically validated. The result indicates the applicability, efficiency, and accuracy of the proposed method.

1 INTRODUCTION

As-built three-dimensional (3D) civil infrastructure modeling, tagged with temporal information for applications such as visual construction simulation can significantly enhance the understanding and communication among various participants involved in different phases (planning, design, construction and maintenance) of an infrastructure life cycle (GSA 2010). Typically, researchers either resort to computer-aided design (CAD) or proprietary code for virtual reality (VR) development to produce 3D models of the constructed facilities (Retik and Shapira 1999; Kamat and Martinez 2001; Al-Hussein et al. 2006), or utilize the range imaging devices such as light detection and ranging scanners (LiDAR) or 3D range cameras to sense the spatial characteristics of the environment and produce a “cloud” of points about the scene of an infrastructure (Jaselskis et al. 2005; Bosche et al. 2009). Both types of modeling methods can achieve high geometric accuracy. However, the CAD/VR based modeling methods tackle 3D modeling of a new object from scratch, requiring sketching the skeleton of the model manually, thereby placing tedious and repetitive demands on the modeler (Dai and Lu 2008). High instrument cost of utilizing range imaging devices sometimes makes the second type of approaches infeasible for small construction projects, where the projected savings hardly justify adopting this technology. In particular, LiDAR scanners are not portable, and require at least two crews to complete site surveying.

With the advances of machine vision and photogrammetry, data from two-dimensional (2D) digital images can be extracted to reconstruct 3D information, which is of great assistance for civil engineers and researchers, for example, to sketch profiles of constructed facilities, measure structural geometric dimensions, and track progress states of an ongoing product. In contrast to CAD/VR- and range imaging-based modeling methods, vision-based modeling methods have the advantages of cost-effectiveness and operational efficiency. It requires the engineer to simply take snapshots of an infrastructure with an off-the-shelf digital camera from different angles. Back in the office, the engineer derives the as-built measurements through post processing of those photos by use of structure from motion (SfM) and photogrammetric algorithms (Dai and Lu 2010).

Computer vision and construction communities have conducted research exploring the ability of visual sensing to recover 3D information from images for use in different domains and applications, such as automated passive recovery of 3D scenes from images and video (Nistér 2004a), unsupervised 3D object recognition and reconstruction (Brown and Lowe 2005), internet photo collection based modeling (Snavely, Seitz, and Szeliski 2008), video-based real time urban 3D modeling (Pollefeys et al. 2008), and four dimensional augmented reality (Golparvar-Fard, Peña-Mora, and Savarese 2009). However, these research efforts principally focus on the scene recovery which contains both background and foreground of an object. For engineering applications, people usually attach more importance to recovering the object itself. A method is needed that will not only extract a point cloud of a complete scene from images, but is also capable of recognizing and detecting the object on top of the scene, therefore obtaining a surface representation of the object.

To address this need, the authors proposed a framework where infrastructure can be reciprocally reconstructed by (1) generating a sparse point cloud of the scene, (2) building up the scene to a dense level, (3) detecting structural members, (4) rendering a 3D view of the structure members, and (5) applying the spatial data covered by regions of the structural members to identify the structure on the generated dense point cloud, and progressively complement or crop the dense point cloud into an object-based as-built model of the structure in place.

The above framework is the final goal. The research in this paper focuses on the first step of recovering a sparse point cloud of the scene using a monocular video camera, which tackles the following three questions that have not been addressed by current existing methods.

- Current reconstruction is usually performed on static and discrete images or photos. The temporal information of video has not been fully exploited, which could be used to substantially reduce the time required for computing the correspondence features between frames.
- Recovery of 3D from video is sporadically investigated in the computer vision community. Most of the research is conducted in ideal laboratory environments. For practical civil engineering settings, real-life constraints have to be taken into account, such as the image blur induced by shakes as the camera moves and videotapes. The practical real-life constraints seriously undermine the chance of a successful 3D reconstruction.
- In addition, recovering the 3D scenes usually targets the visualization-based applications such as monitoring of project progress (Golparvar-Fard, Peña-Mora, and Savarese 2009), and 3D simulation visualization (Dai and Lu 2008). To apply in engineering applications, accuracy of the resulting geometric information is a topic that has not been thoroughly investigated. Dai and Lu (2010) assessed the accuracy for taking geometric measurements of building products, but for their case, a manual interest point selection is required. For automatic 3D reconstruction methods, the resulting accuracy has yet to be systematically assessed.

With these questions, the remainder of this paper reviews the state of research and relevant techniques first. Then, a pragmatic scene recovery method from video using a monocular camera is proposed. Lastly, a case study is demonstrated in which the scene of a reinforced concrete bridge is videotaped, reconstructed, and metrically validated.

2 RESEARCH BACKGROUND

2.1 Image Quality Assessment

In practice, frames extracted from video for spatial data extraction suffer from low quality problems (i.e. blur and noise) if the video is taken by a moving camera. This substantially affects the quality and quantity of the feature points on the extracted frames that can be detected and matched. To resolve this problem, one strategy is to employ a metric that can automatically assess each extracted frame. If the assessed result is above the preset threshold of the metric, the frame is kept and sent to the post-processing step; otherwise, the frame is rejected.

According to Mikolajczyk and Schmid (2005), image color variation due to illumination change or compression of images does not significantly influence feature detection and matching operations. Instead, image blur induced by camera motion or poor focus of the camera lens, will lead to the failure of the image matching operations, or degraded accuracy. For example, as illustrated in Figure 1, two images of a bridge are shot from the same viewpoint with the resolution set at 3888×2592 pixels. The first image is in good quality condition while the second is intentionally distorted by a motion blur. When these two images were used respectively to correlate with the consecutive frame to detect and match feature points, the image pair with the good frame yielded the 2821 correspondence pairs, while the pair with the blurred frame only had 13 correspondence pairs, which are beneath the practical needs to estimate the camera poses.



Figure 1: Comparison of two images of a bridge shot from the same position with one in good quality condition and the other distorted by a motion blur

Image and video quality assessment (QA) algorithms are normally classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) QA algorithms (Wang and Bovik 2006). The FR and RR quality assessment algorithms need access to a “perfect version” or partial information of a “perfect version” of the image or video to evaluate the “distorted version”. The “perfect version” generally comes from a high-quality acquisition device, and the “distorted version” refers to the image or video to be assessed. These two algorithms are not feasible for the proposed method because it is impractical for engineers to prepare a “perfect version” image/video database, requiring both professional quality assessment skills and expensive acquisition devices. Instead, the NR QA algorithms require only the “distorted version” to predict the quality scores (Wang and Bovik 2006), among which there exist a plethora of algorithms that seek to assess the quality of blurred images. However, most of these blur NR QA algorithms are based on evaluating the widths of intensity edges (Marziliano et al. 2002; Chung et al. 2004; Varadarajan and Karam 2008), which may not reflect the quality of the entire image. Chen and Bovik (2009) proposed a multi-resolution decomposition method to extract reliable features of blurred images, but such method requires a process of training sample images, which imposes the extra work on the engineers. Apart from these methods, Crete et al. (2007) developed a simple yet practical no-reference

blur metric based on the discrimination between different levels of blur perceptibility on the same image. For this metric, the complete natural scenes are considered and the resulting scores are simply represented ranging from 0 to 1, respectively the best and the worst quality in terms of blur perception. Therefore, the no-reference blur metric (Crete et al. 2007) fits best for the proposed method to measure the blur effect of the extracted video frames.

2.2 Selection of Key Frames

Unlike static and discrete images or photos, frames extracted from video streams for the estimation of camera motion and object structure may contain excessive information. Rather than process each frame, there should be a way to decide which subsequence of frames to select. A small baseline between two frames may lead to degeneracy of the camera motion estimates, while a large baseline between two frames may cause insufficiency of the feature correspondences, both of which will cause the failure of the subsequent camera pose and scene structure estimations.

To avoid these, key frames should be carefully selected. In the case of initialization for frame sequences, Pollefeys et al. (2002) employed the geometric robust information criterion (GRIC), which was proposed by Torr, Fitzgibbon, and Zisserman (1999), to select the first image pair. The GRIC evaluates which model - homography or fundamental matrix fits better to a set of corresponding feature points in two-view geometry. It guarantees a certain baseline and a large number of initial corresponded feature points. Gibson et al. (2002) proposed a similar method in which three weighted addends are used to make the pairing of frames. The limitations of both methods are, for instance, that a key-frame pairing with a very large baseline is not valued better than a pairing with a baseline that just ensures the fundamental matrix fits better than the homography matrix. Thus, only the degenerate configuration of a pure camera rotation between the key-frame pairing is avoided (Thormählen, Broszio, and Weissenfeld 2004). To ameliorate this situation, Thormählen, Broszio, and Weissenfeld (2004) proposed a criterion based on which the key-frame pairing is selected with the lowest expected estimation error of initial camera motion and object structure. Validated with the ground truth, experiments reveal that Thormählen's method outperforms Pollefeys' and Gibson's methods in estimating camera poses and convergence probability of the bundle adjustment (Thormählen, Broszio, and Weissenfeld 2004).

2.3 Identifying Image Correspondences

Generating point correspondences between image frames requires following the steps of detecting features on frames, describing the features, and matching the corresponding features between two frames. Mikolajczyk et al. (2005) and Moreels and Perona (2007) compared current emerging feature detectors, among which the Hessian-Affine and Harris-Affine obtained the highest scores in terms of consistency of speed, number of features detected, and repeatability of detectors under various condition changes (e.g., viewpoint, scale, light, compression, etc.). Building upon the Hessian-based techniques, Lowe (2004) invented the scale-invariant feature transform (SIFT) and Bay et al. (2008) developed the speeded up robust features (SURF), both of which are robust to image translation, scaling, and rotation. Both SIFT and SURF can serve as the feature point detectors and descriptors. When using the feature points detected for structure from motion, SIFT obtains better accuracy than SURF by a factor of 3 in terms of translation, rotation and reprojection errors (Govender 2009). However, SURF runs around two times faster than SIFT and produces more correct matches per time interval (Bauer, Sunderhauf, and Protzel 2007). The tradeoff between accuracy and computation cost should be made for these two methods. In civil and built environments, structures generally manifest themselves with large scales, and thereby a large quantity of feature correspondences are required to construct a point cloud of these structures. If applying SIFT on such huge amount of features takes a few hours in computing the results, SURF will stand out because the long processing time will be reduced to minutes. As a consequence, the scope of applications for the studied method can be potentially broadened.

Once the feature points on images are detected and described, a matching strategy should be applied to find identical features across these frames. In general, strategies of feature matching include (1) threshold-based: two features are matched if the distance between their descriptors is below a threshold, (2) nearest neighbor: two features are matched if the descriptor of one is nearest to the descriptor of the other and the distance between their descriptors is below a threshold, and (3) nearest neighbor distance ratio: feature A is first nearest to feature B and second nearest to feature C, feature A and feature B are matched if the ratio of feature A and B descriptors' distance over feature A and C descriptors' distance is below a threshold (Mikolajczyk and Schmid 2005). By the threshold-based method, a feature can have several matches but only one of them is correct. With the second and third methods, a feature has only one match and the resulting correct matches are similar (Mikolajczyk and Schmid 2005). However, the nearest neighbor distance ratio matching is computationally more complex. Therefore, this research employs the nearest neighbor matching to perform the matching operations. The nearest neighborhood k-d tree method (Friedman, Bentley, and Finkel 1977; Bentley 1975) will be implemented in the proposed method.

2.4 Camera Pose Estimation

Before using the image pairs to compute spatial positions of the feature points, camera poses in terms of three spatial coordinates and three rotational angles need to be estimated. Also, a good initial estimate of camera poses is vital for ensuring bundle adjustment to achieve successful convergence at a global optimum of the results (Triggs et al. 2000). The camera pose information can be denoted by the matrix multiplication $\mathbf{K}[\mathbf{R} \ \mathbf{t}]$, where \mathbf{K} is a 3×3 matrix containing the intrinsic camera parameters (i.e., the position of the principal point, and the focal lengths expressed in pixel-related units), \mathbf{R} is the 3×3 rotation matrix represented by the camera's three rotational angles, and \mathbf{t} is a three-dimensional vertical vector about the camera translation (Hartley and Zisserman 2004). The intrinsic camera parameters \mathbf{K} can be obtained by camera calibration or by referring to the camera specifications. The rotation and translation matrices \mathbf{R} and \mathbf{t} contain the camera extrinsic parameters, which should be derived by specific algorithms.

To estimate the camera extrinsic parameters, generally the normalized eight point (Hartley 1997), seven point (Hartley and Zisserman 2004), six point (Pizarro, Eustice, and Singh 2003), and five point (Nistér 2004b) algorithms are exploited. The number of eight, seven, six and five is the minimal amount of points required to perform the estimation. Once more points than the minimal amount is provided, the least square adjustment can be applied to work out the optimal solution. Among these algorithms, as tested by Rodehorst, Heinrichs, and Hellwich (2008), the five point algorithm performs best under the influence of artificial noise. Rashidi et al. (2011) validated the accuracy of these algorithms in the practical civil infrastructure environments, resulting in the five point algorithm still outperforming other methods.

In reality, camera lens distortions may cause the image point, which is projected from the object space onto the camera image plane, to shift from its true position to a perturbed position. As a result, two rays of light projected back from two image planes into the object space, may not intersect at a point. This triangulation problem is generally addressed by the Hartley-Sturm algorithm (Hartley and Sturm 1997) or the optimal correction (Kanatani, Sugayak, and Niitsuma 2008) by adjusting the image point positions to minimize the reprojection errors. As tested by Fathi and Brilakis (2011), when applying on an ongoing building reconstruction, the optimal correction performs better than the Hartley-Sturm algorithm in terms of the error minimized and the running time spent. Thereby, the optimal correction is utilized in the proposed method.

2.5 Optimization of Solution

As the features of a scene continue to be revealed and reconstructed from a sequence of video frames, the estimated camera poses and structure positions may suffer from the progressive accumulation of propagated errors induced by preceding calculations. To avoid this, the bundle adjustment is applied on the resulting estimates of camera poses and structure positions obtained in Section 3.4 (Triggs et al. 2000). However, the efficiency of the whole process is decreased if the bundle adjustment is applied. Engels,

Stewénius, and Nistér (2006) disclaimed this concern by providing evidence that a significant amount of bundle adjustment can be performed every time a new frame is added with current computing hardware, even under stringent real-time constraints. Therefore, bundle adjustment is considered in the proposed method, which substantially increases the accuracy of the final result while still maintaining certain computational practicality for the whole process. As a widely used implementation, the Levenberg-Marquardt algorithm based generic sparse bundle adjustment package (Lourakis and Argyros 2009) can be directly adopted for use in this study.

3 PRAGMATIC MONOCULAR SCENE RECOVERY FROM VIDEO

In the first step toward establishing full automation of the modeling process for the proposed framework, a pragmatic method to reconstruct the civil infrastructure scene from video by using a single camera is proposed. Figure 2 shows the detailed workflow of this method.

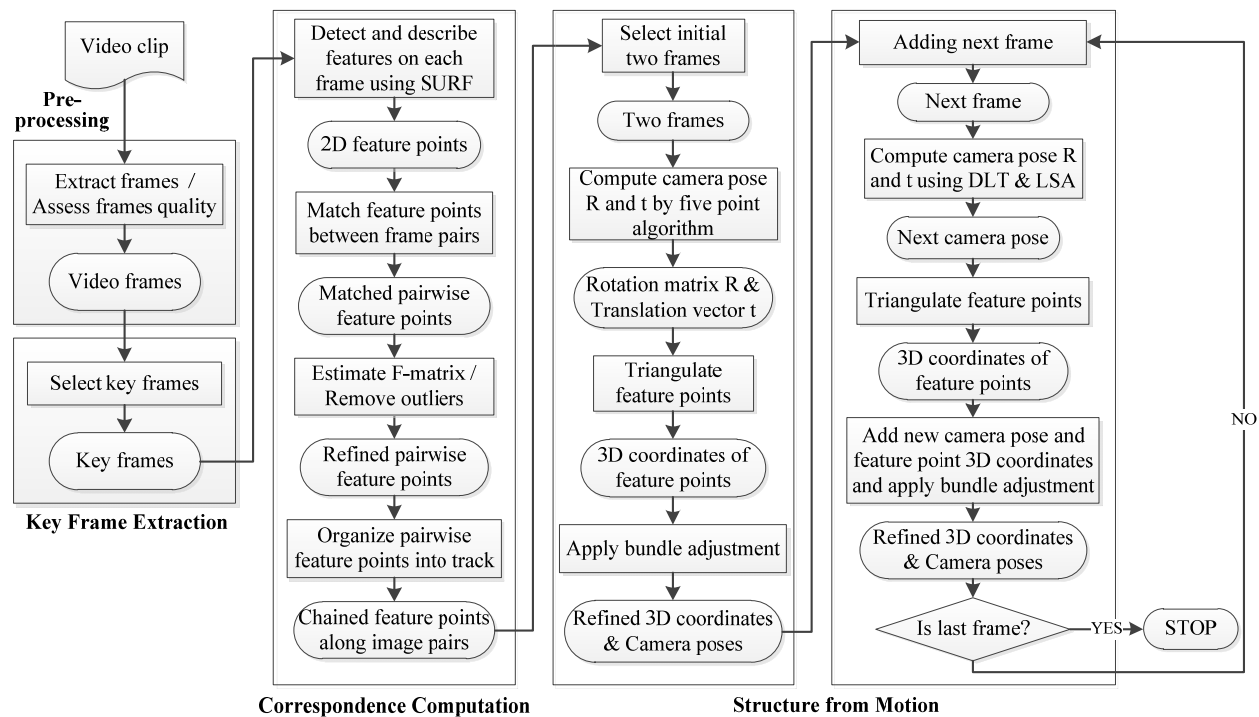


Figure 2: Detailed workflow of proposed monocular scene recovery method

First, frames are extracted from the collected video clip and blurred frames are filtered based on the no-reference blur metric (Crete et al. 2007). Then, the key frames are selected with the lowest expected estimation error of camera poses and feature positions (Thormählen, Broszio, and Weissenfeld 2004). SURF tracker (Bay et al. 2008) is further applied to compute the correspondences of image pairs and chain pairwise the image pairs for later correlation of the camera sequences. Then, the five point algorithm (Nistér 2004b) is used to estimate the initial camera rotation matrix and translation vector, and the feature positions on the first two key frames are triangulated (Kanatani, Sugayak, and Niitsuma 2008). Following that, the bundle adjustment (Lourakis and Argyros 2009) is applied to find the optimal solution of the resulting estimate. Once the first two key frames are processed, the algorithm checks for the existence of new key frames. If there is a new key frame, it will be processed with the established cameras' status to compute the new camera pose and feature positions. Note, from this frame on, instead of the five point algorithm, the camera pose is computed using the traditional direct linear transformation (DLT) method, which requires at least six pairs of feature points to solve the solution (Abdel-Aziz and Karara

1971). In cases that the feature point pairs are more than six, the least square adjustment (LSA) is applied to achieve a more statistically significant and reliable result. Similarly, the 3D coordinates of the new feature points are calculated by applying the triangulation and bundle adjustment. Such sequence will be applied on the next frame and loop until all frames are processed.

4 EXPERIMENTS AND RESULTS

A C# based prototype was implemented to test the validity of the proposed method. A reinforced concrete bridge located on the Interstate-75, in McDonough, GA, was selected as a test bed for the experiment. It is a four-span bridge with three rows of piers and three rectangle columns in each row. A calibrated high-resolution 8-megapixel camera Nikon Coolpix L19 was used to capture the video stream of this bridge. The shooting distance between the camera and the bridge was approximately 30 m. 60 frames were extracted from a 20 sec long video stream taken with a shooting rate of 3 frames per sec. The resolution of these frames is 3264×1840 pixels.

According to Sheikh, Bovik, and de Veciana (2005), an obvious way to measure the quality of an image or video is to solicit opinion from human observers. Thus, 20 random frames with visually satisfactory quality out of those 60 frames were observed and selected to statistically estimate the threshold of the blur metric. Based on the threshold, the blur metric can be incorporated into the automatic process of the proposed method. Applying the blur metric to these 20 frames resulted in the sample mean 0.283, and the sample standard deviation 0.01 of the evaluation scores. The left-sided 95% confidence limit for a normal distribution is then used to determine the statistical estimate of the threshold as $0.283 + 1.64 \times 0.01 = 0.299$. This means that with 95% likelihood, any “satisfied” image would have a score measured by the proposed metric no more than 0.299. It is noted that the sample size is statistically significant to the threshold analysis in consideration of the expected accuracy level being in the order of 0.001 and the relatively small variation on the sample standard deviation of the resulting scores (Hansen, Hurwitz, and Madow 1953). 16 blurred images were filtered. These blurred images were tested with their subsequent images to perform the detection and matching operations, neither of the output feature correspondences were larger than 15, failing to meet the practical quantity required to estimate camera poses (generally over hundred). This verified the validity of the blur metric used and the threshold estimated.

Following the blurred image filtering process, 23 key frames out of the remaining 44 images were selected based on the criterion proposed by Thormählen, Broszio, and Weissenfeld (2004). Figure 3 illustrates eight out of these selected key frames. To validate the correctness of the applied key frame selection method, some of the 21 unselected non-key frames were randomly tested by performing the following motion estimation and bundle adjustment. The result showed that those tested frames either caused the degeneracy of the camera and structure estimation, as visually examined (e.g., Figure 4), or produced larger minimization errors (e.g., $9.23e-3$) of the final bundle adjustment. The error minimized by bundle adjustment with only 23 key frames involved was $6.77e-3$.



Figure 3: Illustrated eight extracted key frames of the bridge

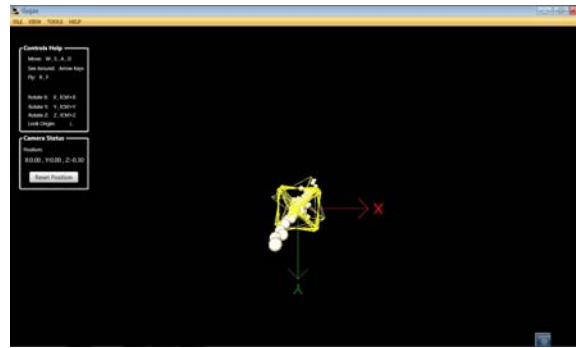


Figure 4: Illustrated example of degenerated camera poses and feature point positions

In parallel to the video stream capturing, a SOKKIA Series 30R reflectorless total station was used to collect data as the ground truth (Figure 5a). As a result, around 2000 points spreading the surface of the bridge were captured and their spatial coordinates were derived. These points were then rendered into a surface model of this bridge to validate the accuracy of the proposed method, as shown in the developed prototype environment (Figure 5b).

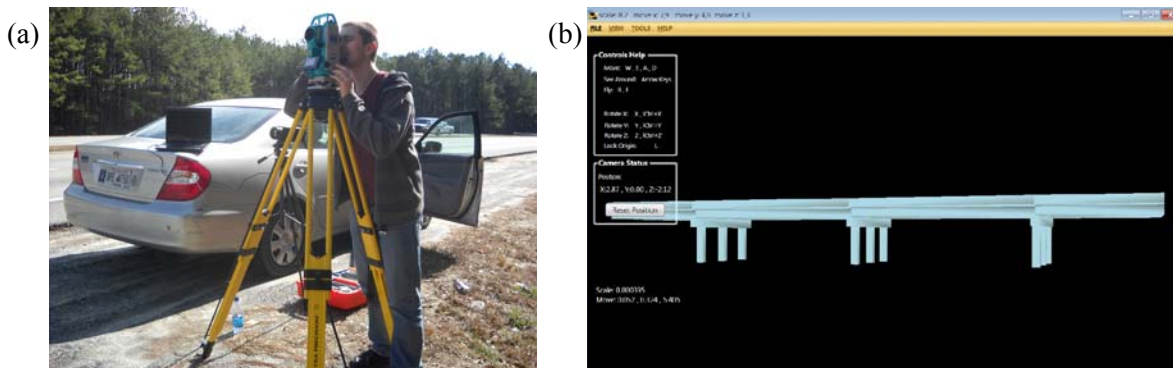


Figure 5: (a) Total station used to collect ground truth data, and (b) actual surface model of the bridge

For each of 22 frame pairs, 250 - 350 features were generally detected and matched before outliers were removed for both SIFT and SURF. After applying the fundamental matrix to remove outliers, there were 170 - 230 remaining features for every frame pair. Tested on one random pair of frames, SIFT and SURF with 64- and 128-dimensional descriptors produced all the correct matches of 98%, while the times consumed for these three algorithms were 4.7, 2.9, and 3.4 sec respectively. According to this, SURF with 64-dimensional descriptors with shortest running time proved to be best candidate for the proposed method. After that, the sparse point cloud and camera trajectory were estimated and visualized in the developed prototype environment (Figure 6a). The camera is denoted by yellow pyramid, and the point is represented by white dot. The total number of 3D points was 1756. It is noteworthy that the running time for processing the 23 frames and producing the point cloud was only 94 sec, with a Dell Vostro 1510 (Intel Core 2 Duo CPU T9300 @2.5 GHz, and RAM 4.0 GB) laptop computer.

The effectiveness of the remaining sequences continued to be validated through evaluating the accuracy of the final resulting point cloud. To conduct this, the point cloud was registered into the coordinate frame where the ground truth model locates using Horn's absolute orientation method (Horn 1987), as shown in Figure 6b. Euclidian distance (error) from a point to surface of ground truth model where this point is supposed to locate, is considered as the metric to measure the accuracy. Denote the i th point's coordinate as (X_i^j, Y_i^j, Z_i^j) , and it is supposed to lie on the j th surface of the ground truth bridge model, as $a_j X + b_j Y + c_j Z + d_j = 0$, the average error of the point cloud can be accordingly calculated by

$$err = \frac{1}{\sum_{j=1}^n m_j} \sum_{j=1}^n \sum_{i=1}^{m_j} \frac{|a_j X_i^j + b_j Y_i^j + c_j Z_i^j + d_j|}{\sqrt{a_j^2 + b_j^2 + c_j^2}}$$

In this equation, m_j is the number of points supposed to belong to the j th surface, and n is the number of the surfaces. 40 points from the frontal surfaces of the bridge were used to evaluate the proposed method (Figure 6b), resulting in an average error of 6.74 cm. The result reveals the proposed method eclipses the method proposed by González-Aguilera and Gómez-Lahoz (2009). In their method, the average error obtained is about 10 cm. The accuracy of the proposed method was also tested under the condition where the bundle adjustment was not applied. This time, the average error yielded was 18.95 cm, which strongly demonstrates the importance of the bundle adjustment as regards to the accuracy of the proposed method.

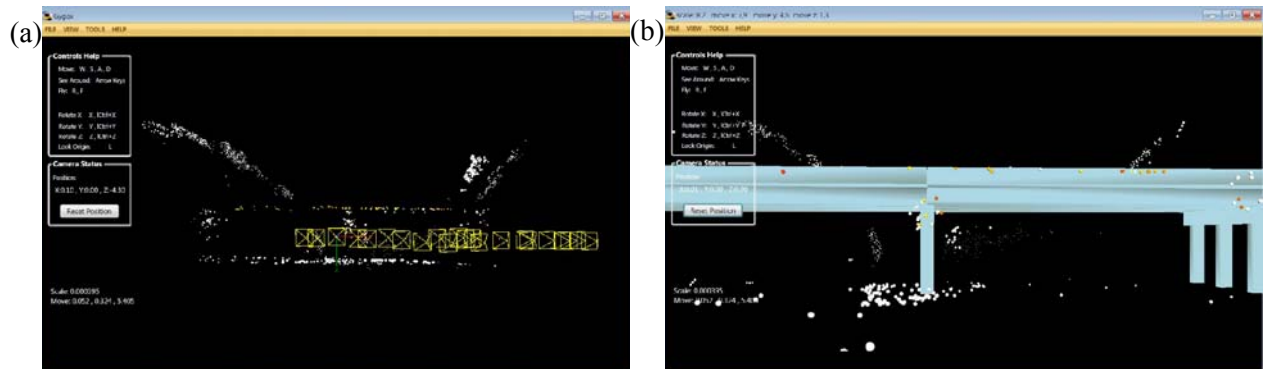


Figure 6: (a) Sparse point cloud of the scene and camera trajectory by the proposed method, and (b) the sparse point cloud registered over the ground truth bridge model for validation

5 CONCLUSIONS AND FUTURE WORK

This paper presented the method of generating the sparse point cloud of a civil infrastructure scene using a single video camera under practical constraints, as the first step toward the establishment of an automatic framework according to which an infrastructure can be reciprocally reconstructed. A case study was demonstrated in which the scene of a reinforced concrete bridge located in McDonough, GA was videotaped, reconstructed, and metrically validated. The result reveals the validity of the proposed method in regard to applicability, efficiency, and accuracy when recovering 3D infrastructure scenes.

The research will be continued in order to seek an effective way of optimizing the performance of the current method. The proposed method will be validated under various settings such as different cameras, different shooting distances, and employing paralleled computation mechanisms. An extension of the ongoing research will use longer video streams to challenge the proposed method in generating “dense” point cloud and employ an adequate number of points to obtain a statistically confident result of accuracy. Existing counterpart methods will be compared with the proposed method. In addition, the remaining steps of the aforementioned framework will continue to be studied and validated.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1031329. The authors gratefully acknowledge NSF’s support. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. The authors are also grateful to the Georgia Tech undergraduate student Craig Burgess, who helped collect the ground truth data for this research.

REFERENCES

- Abdel-Aziz, Y.I., and H. M. Karara. 1971. "Direct Linear Transformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry." In *Proc. of ASP/UI Symp. on Close-Range Photogrammetry*, 1-18. Falls Church, VA: American Society of Photogrammetry.
- Al-Hussein, M., M. A. Niaz, H. Yu, and H. Kim. 2006. "Integrating 3D Visualization and Simulation for Tower Crane Operations on Construction Sites." *Automation in Construction* 15(5): 554-562.
- Bauer, J., N. Sunderhauf, and Protzel, P. 2007. "Comparing Several Implementations of Two Recently Published Feature Detectors." In *Proc. of the Int. Conf. on Intelligent and Autonomous Systems*.
- Bay, H., A. Ess, T. Tuytelaars, and L. V. Gool. 2008. "SURF: Speeded Up Robust Features." *Computer Vision and Image Understanding (CVIU)* 110(3):346-359.
- Bentley, J. L. 1975. "Multidimensional Binary Search Trees Used for Associative Searching." *Communications of the ACM* 18(9):509-517.
- Bosche, F., C. Haas, and B. Akinci. 2009. "Performance of a New Approach for Automated Recognition of Project 3D CAD Model Objects in Site 3D Laser Scans." *Journal of Computing in Civil Engineering* 23(6): 311-318. Reston, VA: ASCE.
- Brown, M., and D.G. Lowe. 2005. "Unsupervised 3D Object Recognition and Reconstruction in Unordered Datasets." In *Proc. of Int. Conf. on 3-D Digital Imaging and Modeling (3DIM 2005)*, 56-63. Piscataway, NJ: IEEE.
- Chen, M.J., and A. C. Bovik. 2009. "No Reference Image Blur Assessment Using Multiscale Gradient." In *Proc. 1st Int. Workshop on Quality of Multimedia Experience*, 70-74. Piscataway, NJ: IEEE.
- Chung, Y., J. Wang, R. Bailey, S. Chen, and S. Chang. 2004. "A Nonparametric Blur Measure Based on Edge Analysis for Image Processing Applications." *IEEE Conference on Cybernetics and Intelligent Systems*, 1, 356-360. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Crete, F., T. Dolmiere, P. Ladret, and M. Nicolas. 2007. "The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric." In *Proceedings of the SPIE*, edited by B. E. Rogowitz, T. N. Pappas, and S. J. Daly, 6492:64920I. Bellingham WA: SPIE.
- Dai, F., and M. Lu. 2008. "Photo-Based 3D Modeling of Construction Resources for Visualization of Operations Simulation: Case of Modeling a Precast Facade." In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 2439-2446. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Dai, F., and M. Lu. 2010. "Assessing the Accuracy of Applying Photogrammetry to take Geometric Measurements on Building Products." *Journal of Construction Engineering and Management* 136(2): 242-250. Reston, VA: ASCE.
- Engels, C., H. Stewénus, and D. Nistér. 2006. "Bundle Adjustment Rules." In *Proceedings of the 2006 Photogrammetric Computer Vision (PCV) Symposium*, edited by W. Forstner, and R. Steffen, 266-271. Istanbul, Turkey: International Society for Photogrammetry and Remote Sensing.
- Fathi, H., and I. Brilakis. 2011. "Automated Sparse 3D Point Cloud Generation of Infrastructure Using its Distinctive Visual Features." *Journal of Advanced Engineering Informatics* 25(4): 760-770.
- Friedman, J. H., J. H. Bentley, and R. A. Finkel. 1977. "An Algorithm for Finding Best Matches in Logarithmic Expected Time." *ACM Transactions on Mathematical Software* 3(3): 209-226.
- Gibson, S., J. Cook, T. Howard, R. Hubbard, and D. Oram. 2002. "Accurate Camera Calibration for Off-Line, Video-Based Augmented Reality." In *Proceeding of International Symposium on Mixed and Augmented Reality (ISMAR 2002)*, 37-46. Piscataway, NJ: IEEE.
- Golparvar-Fard, M., F. Peña-Mora, and S. Savarese. 2009. "Application of D4AR – A 4-Dimensional Augmented Reality Model for Automating Construction Progress Monitoring Data Collection, Processing and Communication." *Journal of Information Technology in Construction* 14: 129-153.
- González-Aguilera, D., and J. Gómez-Lahoz. 2009. "Dimensional Analysis of Bridges from a Single Image." *Journal of Computing in Civil Engineering* 23(4): 319-329. Reston, VA: ASCE.

- Govender, N. 2009. "Evaluation of Feature Detection Algorithms for Structure from Motion." In *Proc. of the 3rd Robotics and Mechatronics Symposium*, 4. Pretoria, South Africa: Council for Scientific and Industrial Research.
- GSA (General Services Administration). 2010. "3D-4D Building Information Modeling." GSA Building Information Modeling Guide Series. Accessed December 12. <http://www.gsa.gov/bim>.
- Hansen, M. H., W. N. Hurwitz, and Madow, W.G. 1953. "Simple random sampling." *Sample Survey Methods & Theory, Volume I Methods and Applications*. NY: John Wiley & Sons., Chap. 4, 126-129.
- Hartley, R. 1997. "In Defense of the Eight-Point Algorithm." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(6):580-593.
- Hartley, R., and P. Sturm. 1997. "Triangulation." *Journal of Computer Vision and Image Understanding* 68(2):146-157.
- Hartley, R., and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge, UK : Cambridge University Press.
- Horn, K. P. B. 1987. "Closed-Form Solution of Absolute Orientation Using Unit Quaternions." *Journal of the Optical Society of America A* 4(4):629-642.
- Jaselskis, E.J., Z. Gao, and R. C. Walters. 2005. "Improving Transportation Projects Using Laser Scanning." *Journal of Construction Engineering and Management* 131(3): 377-384. Reston, VA: ASCE.
- Kamat, V.R., and J. C. Martinez. 2001. "Visualizing Simulated Construction Operations in 3D." *Journal of Computing in Civil Engineering* 15(4):329-337. Reston, VA: ASCE.
- Kanatani, K., Y. Sugayak, and H. Niitsuma. 2008. "Triangulation from Two Views Revisited: Hartley-Sturm vs. Optimal Correction." In *Proc. of the 19th British Machine Vision Conference*, 173-182. Malvern, Worcs, UK: The British Machine Vision Association and Society for Pattern Recognition.
- Lourakis, M.I.A., and A. A. Argyros. 2009. "SBA: A Software Package for Generic Sparse Bundle Adjustment." *ACM Transactions on Mathematical Software (ACM)* 36 (1):1-30.
- Lowe, D. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *Int. J. of Computer Vision* 60(2):91-110.
- Marziliano, P., F. Dufaux, S. Winkler, and T. Ebrahimi, T. 2002. "A No-Reference Perceptual Blur Metric." In *Proc. the International Conference on Image Processing*, 3, 57-60. Piscataway, NJ: IEEE.
- Mikolajczyk, K., and C. Schmid. 2005. "A Performance Evaluation of Local Descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10):1615-1630.
- Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. 2005. "A Comparison of Affine Region Detectors." *Int. J. of Computer Vision* 65(1/2):43-72.
- Moreels, P., and P. Perona. 2007. "Evaluation of Features Detectors and Descriptors based on 3D Objects." *International Journal of Computer Vision* 73(3):263-284.
- Nistér, D. 2004a. "Automatic Passive Recovery of 3D from Images and Video." In *Proc. of the 2nd Int. Symp. on 3D Data Processing, Visualization & Transmission*, 438-445. Washington, DC: IEEE.
- Nistér, D. 2004b. "An Efficient Solution to the Five-Point Relative Pose Problem." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 26(6):756-770.
- Pizarro, O., R. Eustice, and Singh, H. 2003. "Relative pose estimation for instrumented, calibrated platforms." In *Proceedings of 7th Digital Image Computing: Techniques and Applications*, 601-612.
- Pollefeys, M., D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. 2008. "Detailed Real-Time Urban 3D Reconstruction from Video." *International Journal of Computer Vision* 78(2-3):143-167.
- Pollefeys, M., L.V. Gool, M. Vergauwen, K. Cornelis, F. Verbiest, , and J. Tops. (2002). "Video-to-3d." In *Proceedings of Photogrammetric Computer Vision (ISPRS Commission III Symposium)*, International Archive of Photogrammetry and Remote Sensing, 34, 252-258.
- Rashidi, A., F. Dai, I. Brilakis, and P. Vela. 2011. "Comparison of Camera Motion Estimation Methods for 3D Reconstruction Of Infrastructure." In *Proceeding of the 2011 ASCE International Workshop on Computing in Civil Engineering*, edited by Y. Zhu and R. R. Issa, 363-371. Reston, VA: ASCE.

- Retik, A., and A. Shapira. 1999. "VR-based Planning of Construction Site Activities." *Automation in Construction* 8(6):671-680.
- Rodehorst, V., M. Heinrichs, and O. Hellwich. 2008. "Evaluation of Relative Pose Estimation Methods for Multi-Camera Setups." In *International Archives of Photogrammetry and Remote Sensing (ISPRS '08)*, edited by J. Chen, 135-140. Istanbul, Turkey: ISPRS.
- Sheikh, R.H., C. A. Bovik, and G. de Veciana. 2005. "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics." *IEEE Transactions on Image Processing* 14(12):2117-2128.
- Snavely, N., S. Seitz, S., and R. Szeliski. 2008. "Modeling the World from Internet Photo Collections." *International Journal of Computer Vision* 80(2): 189-210.
- Thormählen, T., H. Broszio, and A. Weissenfeld. 2004. "Keyframe Selection for Camera Motion and Structure Estimation from Multiple Views." In *Computer Vision – ECCV 2004*, edited by T. Pajdla and J. Matas, LNCS, 3021:523-535. London, UK: Springer.
- Torr, P., A. Fitzgibbon, and Zisserman. 1999. "The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Images." *International Journal of Computer Vision* 32(1):27-44.
- Triggs, B., P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. 2000. "Bundle Adjustment - A Modern Synthesis." LNCS 1883: 298-375. London, UK: Springer.
- Varadarajan, S., and L. J. Karam. 2008. "An Improved Perception-Based No-Reference Objective Image Sharpness Metric Using Iterative Edge Refinement." In *Proceedings of the 15th IEEE International Conference on Image Processing*, 401-404. Piscataway, NJ: IEEE.
- Wang, Z., and A. C. Bovik. 2006. *Modern Image Quality Assessment, volume 2*. San Rafael, CA: Morgan & Claypool Publishers.

AUTHOR BIOGRAPHIES

FEI DAI is a Postdoctoral Research Fellow in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. He obtained his PhD degree of Construction Engineering and Management from the Hong Kong Polytechnic University. His research interests are advanced information technologies in photogrammetry, computer vision, and simulation and optimization for architecture, engineering, and construction applications (AEC). His email address is feidai@gatech.edu.

ABBAS RASHIDI is a Research Assistant and PhD Student in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. His research interests are reciprocal reconstruction and recognition for modeling of constructed facilities. He is currently working on efficient techniques for 3D reconstruction and 2D element recognition of concrete bridges. His e-mail is rashidi@gatech.edu.

IOANNIS BRILAKIS is an Assistant Professor in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. Before joining Georgia Tech he was an Assistant Professor at the University of Michigan. He obtained his PhD degree in Civil Engineering from the University of Illinois, Urbana Champaign. He is a member of the Editorial Boards of the *ASCE Journal of Construction Engineering and Management*, the *ASCE Journal of Computing in Civil Engineering*, and the *Elsevier Journal of Advanced Engineering Informatics*. His interests are computing and information technologies for the architecture, engineering, construction, and facilities management industries (AEC/FM). His e-mail is brilakis@gatech.edu.

PATRICIO VELA is an Associate Professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. He received his doctorate in Control and Dynamical Systems from the California Institute of Technology. His research interests are computer vision and nonlinear control theory. His e-mail address is pvela@gatech.edu.