Towards Sensor-Aided Multi-View Reconstruction for High Accuracy Applications

Mikhail M. Shashkov, Mauricio Hess-Flores, Shawn Recker, and Kenneth I. Joy
Institute for Data Analysis and Visualization
University of California – Davis
Davis, USA
mmshashkov@ucdavis.edu, mhessf@ucdavis.edu, strecker@ucdavis.edu, kenneth.i.joy@gmail.com

***Abstract*—We present the general idea of a computer vision structure-from-motion framework that makes use of sensor fusion to provide very accurate and efficient multi-view reconstruction results that can capture internal geometry. Given the increased ubiquity and cost-effectiveness of embedding sensors, such as positional sensors, into objects, it has become feasible to fuse such sensor data and camera-acquired data to vastly improve reconstruction quality and enable a number of novel applications for structure-from-motion. Application areas, which require very high accuracy, include medicine, robotics, security, and additive manufacturing (3D printing). Specific examples and initial results are discussed, followed by a discussion on proposed future work.**

Keywords-sensor fusion; embedded sensors; multi-view reconstruction; structure-from-motion; Kinect.

I. INTRODUCTION

In the past few years, there has been a great increase in the amount of sensors that are embedded into every day devices on account of the positive trends in lower costs and miniaturization. For example, consider a modern Android® or iOS® phone whose internal sensors (Global Positioning System (GPS), camera, gyroscope, magnetometer, accelerometer, proximity, audio, and more) drastically outnumber the bigger and less capable cellular phones of prior generations. This trend extends outside of industry and into research where other common sensors, including radar, sonar, LIDAR, infrared, seismic, and magnetic have become utilized more often.

The ubiquity of such sensors and their data creates a *sensor fusion* problem. Sensor fusion involves combining data acquired from different sources in order to provide more accurate or complete information about the sensed target than if these sources were utilized individually. Fusion is non-trivial, and is a very relevant topic today in fields such as computer vision.

In computer vision, one specific instance of sensor fusion is the Red-Green-Blue-Depth (RGB-D) camera, such as the Microsoft Kinect®, which jointly acquires color (RGB) data and depth (D) values for each pixel. The addition of depth freed the Kinect from a certain amount of dependence on analyzing only color to do feature detection, object identification, edge detection, and other fundamental parts of object reconstruction. This boon for research in reconstruction and many other fields culminated in KinectFusion [1], which we describe in the next section. But even the KinectFusion has practical limitations for high-accuracy applications because depth estimates tend to be noisy, and without very accurate filtering, are generally not accurate enough to provide reliable data for up-and-coming applications in medicine, 3D printing and robotics. More traditional methods, like structure-from-motion, space carving and others [2] can be more accurate but are typically less dense. These issues are only exacerbated for additively manufactured objects, which are typically texture-less and mono-colored when produced by current consumer hardware. Figure 1 shows some examples of these objects, including a fully functional ball bearing whose reconstruction would have to be very precise and take into account internal geometry (something the systems discussed cannot do due to occlusions) to maintain functional geometry once reconstructed.

Inspired by the gains achieved from adding depth measurements, we investigate the benefits of using positional sensors to assist in multi-view scene reconstruction. To that end, we present initial results on the development of a generalized framework for 3D scene reconstruction aided by any mix of positional data, such as RGB-D or sonar and photographic imagery. Furthermore, we explore the idea of placing these sensors internally in order to reconstruct internal structure. We show that fusing positional data with traditional images improves the accuracy of camera pose estimation and scene reconstruction, especially when dealing with texture-less or mono-colored objects. This fusion also has the potential to capture internal structure as opposed to standard structure- from-motion approaches. Background and related work is discussed in Section II. Some concrete applications and results will be discussed in Section III. Conclusions and future work will be discussed in Section IV.

Figure 1. Additively manufactured objects that present challenges.

 ISBN: 978-1-61208-342-1

II. BACKGROUND AND RELATED WORK

To our knowledge, we are the first to propose using internally embedded sensors for multi-view reconstruction. We provide a general background on computer vision and contributions towards scene reconstruction in Section II-A, and discuss recent work on fusing sensing technology with imagery, specifically RGB-D cameras, in Section II-B. We will also discuss recent work on the imaging of internal geometry in Section II-C.

A. General Background of Scene Reconstruction

The broad field of computer vision includes important sub-fields such as object detection, tracking, and the multi-view reconstruction of scenes. The goal of multi-view scene reconstruction is to extract a 3D point cloud representing a scene when given multiple views (such as photographs) of the scene. Detailed analysis and comparisons between methods are available in the literature [2]. Most of these methods seek to create correspondences between views, usually by detecting features and tracking them from view to view. One of the main algorithms used to do this is Scale-Invariant Feature Transform (SIFT) [3]. For an excellent overview of many classical vision algorithms, the reader is referred to Hartley and Zisserman [4].

One drawback of current computer vision methods is that many are based on the mathematical optimization of initial parameter estimates to achieve accurate results. Though such optimization is provably necessary, such as in the case of the well-known *bundle adjustment* [5] in structure-from-motion, the final accuracy is simply not enough for applications that require an extreme amount of accuracy. Furthermore, the density of these reconstructions often leaves something to be desired.

B. RGB-D Cameras

To alleviate the density problem, there has been interest in utilizing depth sensing technologies for object reconstruction for a long time [6], but it is only recently that the technology has become very affordable and easy to use with the release of the Microsoft Kinect® in late 2010. With it, came a plethora of reconstructions of people [7] and indoor environments [8]. One of the biggest successes is KinectFusion, which fuses depth data and RGB data from a movable Kinect in real time to create a dense scene reconstruction as the user moves through the scene. Given its ubiquity and success, we will further detail the KinectFusion algorithm [1][9], since this is the main algorithm we want to challenge as far as reconstruction density and accuracy for our intended applications. The main goal of KinectFusion is to fuse depth data acquired from a Kinect sensor into a single, global surface model of the viewed scene, in real-time. Additionally, 6DOF sensor pose is simultaneously obtained by tracking the live depth frame relative to the global model using a coarse-to-fine Iterative Closest Point (ICP) algorithm [20].

The KinectFusion algorithm can be considered an upgrade to previous 'monocular Simultaneous Localization and Mapping (SLAM)' systems [21], the most successful being the Parallel Tracking and Mapping (PTAM) system [10]. The main drawback of those systems is that they are optimized for efficient camera tracking, but produced only sparse point cloud models. Even in novel systems, which combine PTAM's camera tracking capability with dense surface reconstruction methods (such as described in [1]) in order to enable better occlusion prediction and surface interaction [11][12], dense scene reconstruction in real-time remains a challenge. Results are still highly dependent on factors such as camera motion and scene illumination.

KinectFusion has been proven to work well for situations with a dynamic element involved: either the objects in the scene or the camera itself is moving. In our applications, we're more interested in acquiring a very high level of detail, even from a completely static setup. For instance, the KinectFusion algorithm relies on bilateral filtering on the initial depth maps, for noise removal. Though very helpful towards the original algorithm, such smoothing must be further analyzed in our framework, since it reduces noise but effectively also smooths sharp contrasts and levels of detail. Also, though there are proven advantages of tracking against the growing full surface model with respect to frame-to-frame tracking, there is still likely to be drift over long sequences, which will ultimately affect accuracy. Our intended use of ground-truth information effectively eliminates drift, aiding in more-accurate pose estimation and hence scene reconstruction. Furthermore, our framework can fuse any source of positional information, such as embedded internal sensors, and by virtue potentially capture geometry that is not visible to the naked eye. In the next section, work on viewing such "hidden" geometry is discussed.

C. Internal Imaging

Research in internal imaging has been existent for a long time and has resulted in tremendous advances in both medicine and security. Breakthroughs in these domains have largely been a result of physics and biology research. For example, magnets are used to image the gastrointestinal track, radio waves produced by excited hydrogen atoms are used to image the brain (Magnetic Resonance Imaging), and x-rays (Computerized Axial Tomography) and small amounts of radioactive substances (Positron Emission Tomography) are used for tomography. All of these procedures were revolutionary and are now routine. Similar techniques and technologies have recently been used in tandem with computer vision to enhance security. For example, Taddei and Sequeira demonstrated that x-ray tomography equipment could be calibrated using Automatic Pose Estimation (APE) on the silhouettes of shapes [13].

We are also motivated by structural health monitoring, which is concerned with using embedded sensor networks to evaluate structures such as buildings and bridges and provide feedback on cracks, torsion, and other instabilities. For a great overview see Tignola et al. [14]. We believe the sensors and technology used in structural health monitoring will eventually be miniaturized and can thus be expanded upon to be utilized for geometric reconstruction on much smaller objects than bridges. This has motivated us to begin our preliminary exploration of using such internal sensor

 ISBN: 978-1-61208-342-1

networks and fusing them with imagery to improve current reconstruction approaches.

III. SENSOR-AIDED RECONSTRUCTION

In order to provide a valid comparison point for our approach, the first thing we did was perform our own reconstructions using structure-from-motion, space carving, and KinectFusion using a new dataset. For these reconstructions, we used an additively manufactured chess piece and Utah teapot, both in red polylactic acid (PLA) plastic via the Makerbot Replicator 2®. Our motivation in creating this new dataset using 3D printed objects was *a)* we have ground-truth knowledge about the correct geometry *b)* we intend to show the potential benefits of embedding sensors as part of the manufacturing process and *c)* it allows for the object to be materialized by anyone who wishes to try their own physical approach (for example, KinectFusion).

Our results, shown in Figure 2, demonstrate the various problems with these standard approaches for the objects in (a) and (e). Structure-from-motion results, (b) and (f), are meshed reconstructions retrieved from running Visual SfM [15][16] and using Patch-based Multi-view Stereo (PMVS) [17] to densify. While this approach does a decent job of accurately capturing some important details like the crown on the queen and the handle and spout of the teapot, it is clear that the reconstruction is full of holes and not dense enough. It is important to note that the lack of texture and color variance is one of the major problems for structure-from-motion since it largely depends on the presence of lots of unique features for tracking. Another common approach that doesn't depend on texture or color is space carving [18]. In images (c) and (g), you can see that although it does a great job of creating a dense, water-tight, reconstruction by virtue of the approach, it is not accurate enough to capture the sharp tips of the crown and none of the spout or handle of the teapot. Similarly, KinectFusion [1], (d) and (h), creates wonderfully dense objects but fails to capture small details due to the smallness of the objects, inherent noise and hardware limitations. It is also important to note that although these methods will yield better results for larger objects, the results are only aesthetically pleasing and not actually precise, hence why small features will be missed.

In light of these results, we developed an alternative reconstruction pipeline which couples positional information with structure-from-motion. In general, accurate structure-from-motion based reconstruction typically relies on accurate *feature tracking* [4]. A feature track is a set of pixel positions representing a scene point tracked over a set of images. Given a 3D position computed from multi-view stereo, its *reprojection error* with respect to its corresponding feature track is the only valid metric to assess error, in the absence of ground truth. Highly inaccurate individual track positions adversely affect subsequent camera pose estimation and structure computation, as well as bundle adjustment. Such inaccuracies can be improved upon by including external sensor information, such as positional information, into solving for scene reconstruction. The advantage of counting with embedded positional information inside an object is that it avoids having to compute accurate feature tracks in order to perform camera parameter and structure computation.

A diagram of our pipeline is shown in Figure 3. The process begins by collecting both the positional sensor data and image data. Provided with a mechanism for locating the positional sensor in each image, the accurate position information is used to perform camera pose estimation. This leads to accurate camera rotation and translation measurements and is void of the inaccuracies present when using feature tracks to estimate camera pose. Feature

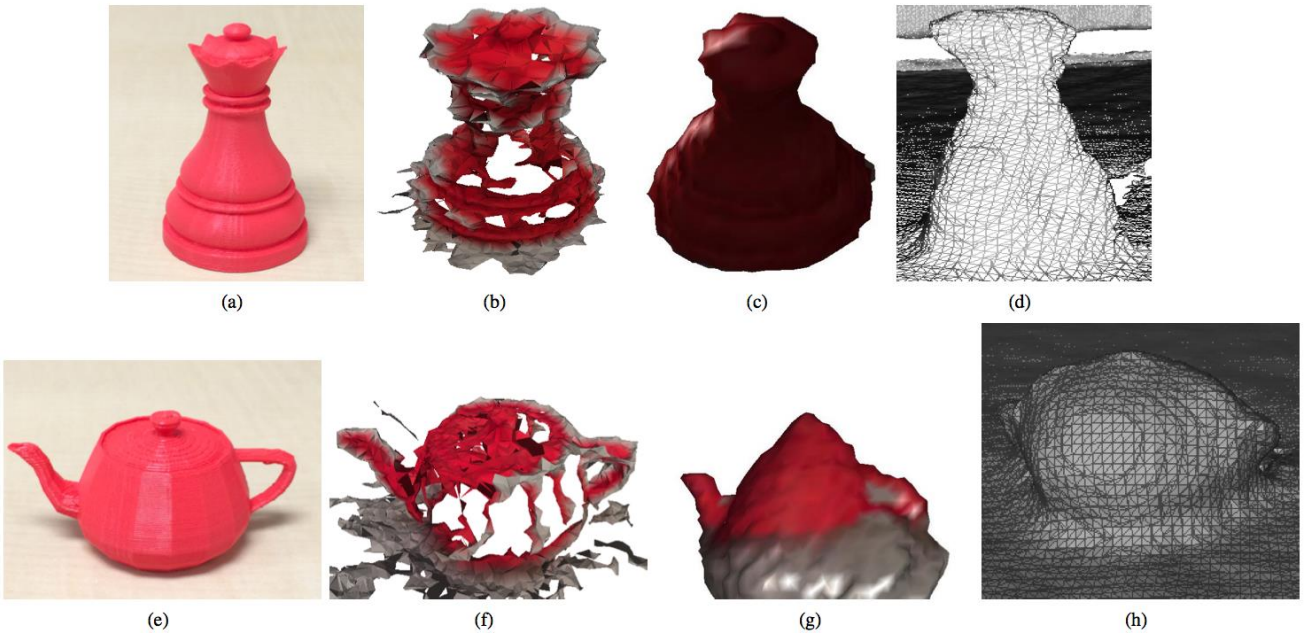


Figure 2. Input (a), structure-from-motion (b), space carving (c), and KinectFusion (d) reconstructions for a 3D-printed red chess piece. The same is shown for the "Utah teapot" in (e) - (h).

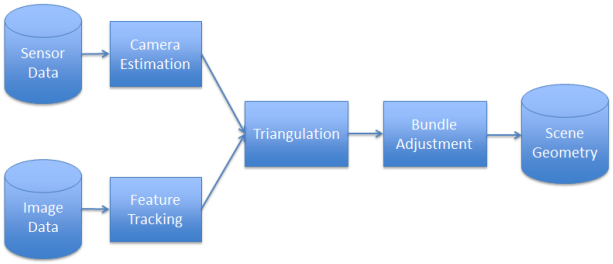


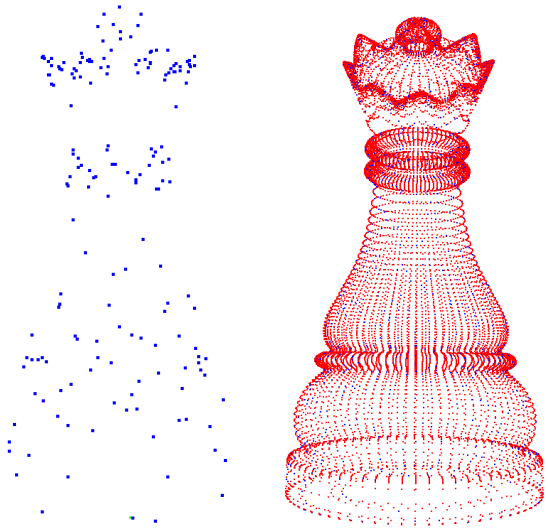
Figure 3. Our proposed reconstruction pipeline that utilizes sensors.

tracking is performed on the image data and is combined with the accurate camera data to perform triangulation of the scenes 3D structure. Errors in the feature tracking stage are manifested as inaccurate scene points and bundle adjustment is used to optimize the reprojection error of the given structure and camera parameters. After bundle adjustment has been performed, the scene geometry can be stored and manipulated using standard modeling techniques.

While our pipeline is only a work-in-progress, initial results show the positive effect of embedded positional sensors on reconstruction. We successfully performed a simulated reconstruction using synthetic data from the chess piece's geometrical definition (a .obj file) in order to sanity-check the camera estimation portion of our pipeline. To simulate surface-level embedded sensors we chose 185 random vertices from the definition file, whose locations appear in Figure 4a. By creating 10 randomly placed synthetic cameras (not pictured) and reprojecting the sensor locations into the synthetic image plane of each camera, we created feature tracks for each sensor. Using these feature tracks and the corresponding ground-truth locations of the sensors, we performed camera pose estimation using the Efficient Perspective-n-point (EPNP) algorithm [19]. Using feature tracks for all 18504 ground-truth vertices and using our computed cameras to triangulate we were able to achieve a reconstruction with essentially zero reprojection error (see Figure 4b). While this result is expected given that we have perfect feature tracks, we have shown that the camera pose estimation section of our pipeline has been implemented correctly and embedded sensors can be used for nearly perfect camera pose estimation. To complete our work, we would use additional feature tracks derived from SIFT-analyzed photography of an object with real surface-level embedded sensors, we discuss how to do so and the implications in the next section.

IV. Conclusion and Future Work

This paper presented the general idea of using sensor fusion as a strong tool for improving accuracy in computer vision structure-from-motion with the end goal of enabling high accuracy applications. Concrete results were shown for synthetic data, where a simulated object with surface-level position sensors was used to very accurately estimate the set of cameras viewing the object. Given these initial results, we believe the future is bright with regards to fusing sensor measurements for improved multi-view reconstruction, which is the focus of our ongoing work.



(a) (b)

Figure 4. (a) Simulated surface-level embedded sensor locations. (b) A reconstruction using synthetically perfect feature tracks but computed cameras confirms nearly perfect camera pose estimation.

We have identified a number of uses in potential applications. One is the additive manufacturing process. By embedding positional sensors as part of the 3D-printing process, a whole host of opportunities open up. First, you can monitor and analyze the object during the printing process and verify key geometric qualities, such as distances or angles. Secondly, if the sensors are miniaturized to a sufficient degree and placed very densely, it becomes unnecessary to even use structure-from-motion or other techniques since a meshed point cloud of sensor locations can be used as a reconstruction by itself (see Figure 4b). Third, a designer could manipulate the printed object with real world tools, such as chisels and saws, and be able to "scan" the object back into virtual space. A similar process already occurs in structural health monitoring where sensors are mixed with concrete; it is our belief that is only a matter of time before the technology is miniaturized enough for small scale objects and additive manufacturing.

Furthermore, at that miniaturized scale, we can expand the concept of structural health monitoring to medical applications and devices. By embedding sensors in artificial human parts, such as hearts and prosthetics, we enable the medical community to non-invasively monitor any defects that may occur by periodically reconstructing the object and analyzing it.

We strongly believe that miniaturized positional sensors are achievable with some combination of modern technologies such as ultrasound, magnets, and piezoelectrics. Simpler, surface-level sensors requiring manual effort could be created with highly reflective targets or glow-in-the-dark plastic/stickers for easy 2D localization by hand or automated process. Our future work will focus on using prototyped positional sensors to proof-of-concept our approach and its revolutionary applications.

 ISBN: 978-1-61208-342-1

ACKNOWLEDGMENT

This work is supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04- 94AL85000. We give special thanks to our colleagues at the Institute for Data Analysis and Visualization for the useful discussions and support.

REFERENCES

[1] S. Izadi et al. "Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera," in Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, ser. UIST '11. New York, NY, USA: ACM, 2011, pp. 559–568. [Online]. Available: http://doi.acm.org/10.1145/2047196.2047270 [retrieved: April, 2014]

[2] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE Computer Society, 2006, pp. 519–528.

[3] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal On Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[4] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, 2004.

[5] M. Lourakis and A. Argyros, "The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg-Marquardt algorithm," Institute of Computer Science - FORTH, Heraklion, Crete, Greece, Tech. Rep. 340, August 2000.

[6] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," Image Vision Comput., vol. 10, no. 3, pp. 145–155, Apr. 1992. [Online]. Available: http://dx.doi.org/10.1016/0262- 8856(92)90066- C

[7] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," Visualization and Computer Graphics, IEEE Transactions on, vol. 18, no. 4, pp. 643–650, 2012.

[8] A. Majdi, M. C. Bakkay, and E. Zagrouba, "3d modeling of indoor environments using kinect sensor," in Image Information Processing (ICIIP), 2013 IEEE Second International Conference on, 2013, pp. 67–72

[9] R. A. Newcombe et al. KinectFusion: Real-Time Dense Surface Mapping and Tracking, in IEEE ISMAR, IEEE, October 2011.

[10] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR), 2007.

[11] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[12] J. Stuehmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In Proceedings of the DAGM Symposium on Pattern Recognition, 2010.

[13] P. Taddei and V. Sequeira, "X-ray and 3d data fusion for 3d re-construction of closed receptacle contents," in 3DV-Conference, 2013 International Conference on, 2013, pp. 231–238.

[14] D. Tignola, S. Vito, G. Fattoruso, F. D'Aversa, and G. Francia, "A wireless sensor network architecture for structural health monitoring," in Sensors and Microsystems, ser. Lecture Notes in Electrical Engineering, C. Di Natale, V. Ferrari, A. Ponzoni, G. Sberveglieri, and M. Ferrari, Eds. Springer International Publishing, 2014, vol. 268, pp. 397–400. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-00684-0 76 [retrieved: April, 2014]

[15] C. Wu, S. Agarwal, B. Curless, and S. Seitz, "Multicore bundle adjustment," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 3057–3064.

[16] C. Wu, "Towards linear-time incremental structure from motion," in 3DV-Conference, 2013 International Conference on, 2013, pp. 127–134.

[17] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," in IEEE Conference on Computer Vision and Pattern Recognition, June 2007, pp. 1–8.

[18] K. Kutulakos and S. Seitz, "A theory of shape by space carving," International Journal of Computer Vision, vol. 38, no. 3, pp. 199-218, 2000. [Online]. Available: http://dx.doi.org/10.1023/A%3A1008191222954 [retrieved: April, 2014]

[19] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. 2009. EPnP: An Accurate O(n) Solution to the PnP Problem. Int. J. Comput. Vision 81, 2 (February 2009), 155-166. DOI=10.1007/s11263-008-0152-6 http://dx.doi.org/10.1007/s11263-008-0152-6

[20] Besl, P.J.; McKay, Neil D., "A method for registration of 3-D shapes," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.14, no.2, pp.239,256, Feb 1992 doi: 10.1109/34.121791 [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=121791&isnumber=3469 [retrieved: April, 2014]

[21] Newcombe, Richard A.; Lovegrove, S.J.; Davison, A.J., "DTAM: Dense tracking and mapping in real-time," Computer Vision (ICCV), 2011 IEEE International Conference on , vol., no., pp.2320,2327, 6-13 Nov. 2011 doi: 10.1109/ICCV.2011.6126513 [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6126513&isnumber=6126217 [retrieved: April, 2014]

 ISBN: 978-1-61208-342-1