Improved Outdoor Augmented Reality through “Globalization”

Chris Sweeney PhD Advisors: Tobias Höllerer & Matthew Turk

Department of Computer Science
University of California, Santa Barbara
{cmsweeney, holl, mturk}@cs.ucsb.edu

ABSTRACT

Despite the major interest in live tracking and mapping (e.g., SLAM), the field of augmented reality has yet to truly make use of the rich data provided from large-scale reconstructions generated by structure from motion. This dissertation focuses on extensible tracking and mapping for large-scale reconstructions that enables SfM and SLAM to operate cooperatively to mutually enhance the performance. We describe a multi-user, collaborative augmented reality system that will collectively extend and enhance reconstructions of urban environments at city-scales. Contrary to current outdoor augmented reality systems, this system is capable of continuous tracking through areas previously modeled as well as new, undiscovered areas. Further, we describe a new process called *globalization* that propagates new visual information back to the global model. Globalization allows for continuous updating of the 3D models with visual data from live users, providing data to fill coverage gaps that are common in 3D reconstructions and to provide the most current view of an environment as it changes over time. The proposed research is a crucial step toward enabling users to augment urban environments with location-specific information at any location in the world for a truly global augmented reality.

1 INTRODUCTION

The rapid growth of online photo collections has allowed for efficient 3D reconstruction of massive datasets efficiently [1, 17]. Once built, these massive models are treated as static and can only incorporate new data with immense effort, often requiring a non-linear least squares minimization of reprojection errors (called bundle adjustment) on the entire set of 3D points and cameras. Further, these data sources often have limited viewpoints: Flickr images are frequently clustered around landmarks, and Street View images are largely limited to roads. As a result, these data sources can have coverage gaps in areas that have not yet been densely imaged, leaving much to be desired.

Emerging technologies such as Google Glass shed light on how these models may be used to quickly locate where a user is in the world so their viewpoint can be seamlessly augmented with useful information about their surroundings. This augmentation relies on live tracking as the user moves about the environment so that the user’s position and orientation are correctly registered. While systems such as PTAM [8] are capable of robust visual tracking, they are predominantly suitable for small indoor scenes. On the other hand, systems that use model-based tracking are ideal for tracking with SfM data; however, these systems are limited by the coverage and detail of the model. A more natural system would allow the user to discover new areas with uninhibited motion. Extensible tracking was a major focus of SLAM research prior to the work of PTAM [3, 6, 14], which subsumes the task of extensible tracking for small areas. However, significant challenges remain for extensible tracking with large-scale, outdoor models.

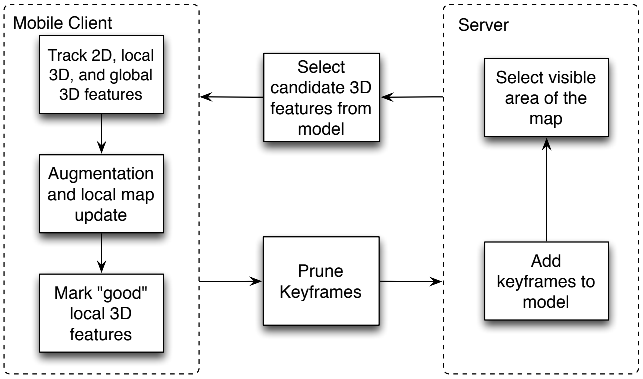
Mobile Client
Track 2D, local 3D, and global 3D features
Augmentation and local map update
Mark "good" local 3D features
Select candidate 3D features from model
Prune Keyframes
Server
Select visible area of the map
Add keyframes to model

Figure 1: Overview of the proposed system that localizes from a previously built large-scale model then provides continuous tracking while actively updating the large-scale model. The middle layer is performed as necessary to provide optimal SLAM and SfM performance.

Additionally, SLAM systems make no effort to utilize the acquired visual data after the augmented reality task is completed. As such, current SfM and SLAM systems operate independently instead of cooperatively, despite the commonalities between them. Making full use of visual data acquired during augmentation would enable crowd-sourcing a global 3D model of urban environments. Crowd-sourcing has several major benefits over manual data acquisition including cost, speed, and coverage.

The focus of this dissertation is to create an agreement between large-scale SfM and SLAM so that large-scale SfM reconstructions can be used to enhance SLAM and vice versa. Based on this agreement, we propose a real-time system that utilizes live data from many AR users to crowd-source the task of reconstructing 3D models for outdoor scenes. The system we describe is capable of localizing from a large-scale model and uses model data in addition to newly acquired visual data to perform live tracking outdoors. Ventura and Hollerer [19] describe a system that performs online tracking similar to PTAM after localizing from a SfM model built offline; however, their approach is currently limited to relatively small outdoor scenes. Further, the proposed system makes use of data acquired during live tracking by broadcasting information about the local map generated from SLAM back to the global model through a new process called *globalization*. By utilizing all data that each user generates during the online tracking and mapping stage, we are able to efficiently fill in holes and extend the 3D model that had previously been generated offline. Globalization is a key element of this dissertation, and we believe it is essential for the goal of making augmented reality globally accessible.

2 PROJECT OVERVIEW

The proposed dissertation is centered around a client-server system that is able to provide stable, continuous tracking for outdoor AR users for long periods of time. The client (e.g., mobile device or head mounted display) is largely responsible for SLAM tasks while the server is responsible for SfM tasks; they are each able to operate independently, but each is greatly enhanced by interacting with the other. Our goal of providing a seamless client-server relationship for outdoor AR requires a seamless relationship between the SLAM and SfM tasks. Indeed, this is the framework in which we approach the problem. Note that we assume a large-scale reconstruction has been acquired using standard SfM techniques and thus a major focus of this dissertation is how to best utilize and enhance that reconstruction as new data is acquired from SLAM. We pose the framework in such a way that each main component of our system is generous to the other: the SfM server puts great effort into providing relevant visual data for continuous tracking, and the SLAM client is able to intelligently provide the SfM server with new visual data that will enhance its model.

2.1 Tracking With or Without Global Context

Our tracking system utilizes typical SLAM in conjunction with model-based tracking. First, we perform localization by querying the server for an absolute camera pose. While localization can be performed on a large-scale model in nearly real-time [7, 11, 12, 15, 21], wireless bandwidth can be unpredictable. To mitigate the delay, we begin visual tracking using the initialization-free system of our previous work [5] so that the user's motion is never constrained. This particular tracking system is well suited for outdoors because of its ability to track through rotation-only motion[1]. We update the poses of any acquired keyframes with the absolute pose when it is available. In this sense, our tracking system is not dependent on localizing to the global model – localization enhances the AR experience with global context (and thus, global augmentation) but tracking degenerates to a simple outdoor SLAM system that is capable of local augmentations when an absolute pose is not available.

During tracking, we label three types of keypoints: 3D features from the global model, 3D features from the local model (ones that are not yet registered in the global model), and 2D features. Tracking solely the first type of keypoint is pure "model-based tracking," while tracking the second and third types (but not the first) is what most PTAM-like systems do. In order to maintain real-time performance, we take a distributed matching approach for 3D features inspired by Lim et al. [13]. Once a 2D feature has been tracked through a significant number of frames we label it as a "good" feature for which we will attempt to recover 3D information. Each frame we process a constant number of "good" features to amortize the cost of expensive 2D-3D matching over several frames. This is done by searching through a set of likely visible 3D features the server has provided the client based on its absolute pose. If a good 2D-3D match is not found, the 2D feature is triangulated with standard SLAM techniques and labeled as a 3D feature from the local model. Eventually, this local 3D feature will be added to the global model through globalization. This process is discussed in detail in the next section.

2.2 Updating the Global Model with Globalization

We take a similar approach to our "lazy" strategy for 2D-3D matching for updating the global model on the server. As online tracking proceeds we limit the number of keyframes that are kept by the local model to maintain a manageable local map size. For now, we choose the simple strategy of removing the oldest keyframes from the local map. For a model M and a set of keyframes K, where the $||*||$ operator denotes the number of 3D points, we choose a subset of keyframes K^* subject to $\arg\max_{K^*}(||K^*|| - ||M \cap K^*||)$. This approach favors adding as much new 3D information to the global model as possible, and new keyframes can easily be added with standard incremental SfM techniques on the relevant subset of the global model, thus avoiding global bundle adjustment.

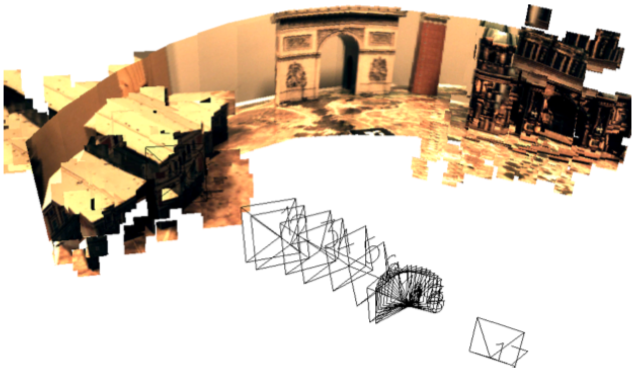

Figure 2: Our system described in [5] alternates actings as a SLAM system and a panorama mapper depending on the current camera motion to continue tracking through movements that would otherwise break either system individually. This results in a combination of 3D structure and panoramic maps to represent the environment, as shown here.

We term this process "globalization." Whereas localization queries information about the current camera location from a 3D model, globalization adds information to a 3D model based on the current camera location (or set of locations). We are able to avoid issues with loop closure during tracking with globalization, as features from the beginning of the loop either existed previously in the global model or are added to the global model during tracking. Loop closure then is subsumed by simple 2D-3D matching. Globalization is able to trivially extend (e.g., add previously un-mapped areas) and enhance (e.g., fill holes) the global model.

As new visual data is added to the global model via globalization, the server also provides the client with new 3D data that is likely visible based on the current camera pose. This data is formed as a set of images chosen to provide maximal visual coverage. In this way, the data provided to the client is comparable to keyframes, so localizing to this new data is simply a matter of keyframe-based recovery. We assume that when the client sends new visual data to the server, it is a good indication that the client is exploring an area where the 3D data it contains locally are no longer visible. If this is not the case (i.e., the 3D features are still visible) then the recently added visual data will only add detail to the model that is currently being viewed. This is a case of loop closure (with hole filling), as discussed previously. By continuously providing the client with 3D data it is likely to see, we allow users to move continuously through large areas without interrupted tracking.

3 Completed Work

In this section, we summarize our relevant completed work. Section 4 will propose future work building upon our existing work described here. We have already begun implementing an open-source computer vision library that provides a common interface for SLAM and SfM techniques so that parts of each pipeline are more easily tuned and adapted than with other currently available software. We feel this is an important contribution to the research community moving forward, as testing the various individual components of SfM and SLAM systems becomes increasingly studied.

[1] PTAM-like systems rely on a minimum motion parallax during initialization that is proportional to the depth of triangulated 3D points. For outdoor scenes this depth can be very large, making the baseline required for initialization impractical and causing tracking to fail.

3.1 Real-time Tracking and Mapping

We presented the first real-time tracking and mapping system that handles both general and rotation-only motion [5] at last year's ISMAR, receiving the "Best Paper Award" for our work. Previously, SLAM based systems were limited to either translational motion or rotation-only motion. As a result, tracking would be lost in many circumstances where the camera motion did not match the particular tracking strategy. For a given pair of successive images, we use a model selection criterion to choose the motion model that can most completely describe the camera motion (translational or rotation-only). This allows us to maintain tracking through motion that would break either a typical SLAM system or a panorama tracker. This is a significant result for outdoor SLAM, which usually relies on panorama tracking because of the relatively far distance of feature points. We recently extended this system to support model-based tracking which resulted in significant tracking improvements (currently in submission to TVGC).

3.2 Pose with Known Vertical Direction

We have derived simple and efficient formulations for determining absolute camera pose from two points and relative camera pose from three points when a vertical direction is known. Preliminary experiments have indicated that our methods are faster, more accurate, and more numerically stable compared to existing methods. This efficient formulation will help speed up tracking on the mobile client where internal hardware is able to provide a global vertical vector. This work will be submitted to CVPR 2014.

4 Future Work

Our completed and ongoing work has given us encouraging results. The problem remains that large-scale SfM has been underutilized in the field of augmented reality. Significant work is required to change this situation.

4.1 Investigation of Binary Descriptors for SfM

Most current SfM pipelines use SIFT descriptors because of their exceptional robustness. Recent work on binary descriptors has led to efficient new descriptors with comparable matching performance to SIFT [2, 10] at a fraction of the cost. Binary features have yet to be thoroughly evaluated for SfM but provide a promising avenue for future research. Such binary features could be used directly for tracking in SLAM, making 2D-3D matching and insertion to the global model (via globalization) much more simple and efficient. We would like to evaluate several descriptors, including SIFT, DAISY, BRISK, and FREAK, to determine their viability for use in large-scale SfM and localization.

4.2 Preemptive Image Matching

Wu [20] describes a preemptive test to quickly determine whether an image pair will be a good match. They note that the feature matches between the first k features with the largest scale provide a good indicator for the inlier ratio of all features in the image pair. However, their current implementation relies on setting the parameter k by hand. By using extreme value theory methods presented by Fragoso and Turk [4], we can adaptively and automatically choose the parameter k based on the distribution of nearest neighbor distances. This will enable a more robust criterion for preemptive matching, resulting in fewer missed image matches with no additional cost. Additionally, we are interested in exploring the use of vocabulary trees for image matching (based on [1]) using the largest-scale features to increase the speed of matching and localization.

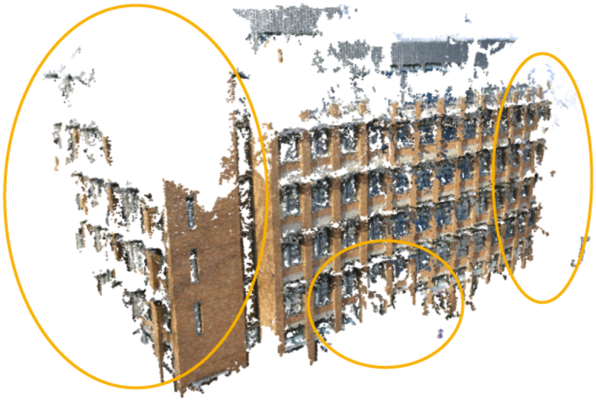

Figure 3: Incomplete sections of a reconstruction (circled) can easily be recovered with globalization (image from [18]).

4.3 Globalization

A critical aspect of this system is the ability to continuously and seamlessly contribute new data to a global model. We are unaware of any previous work that updates a large-scale model continuously through methods similar to globalization. Determining the proper criteria for adding local keyframes to a global model is a major area for exploration going forward. The keyframe pruning described in Sec. 2.2 prioritizes adding as much new 3D information to the model as possible. To limit contamination of the model with "bad" 3D points (which is known to happen for incremental structure from motion), we will explore using image triplets which have been proven to increase robustness [9] for 3D reconstructions. Other visual cues such as vanishing point detection have been shown to increase robustness, as they add additional constraints to ease the work of bundle adjustment [16]. Limiting the cost of expensive bundle adjustment is increasingly important as the global model becomes more dense with new visual information added over time.

5 Research Outcomes

The results of this research will enable several exciting possibilities. Real-time access to 3D data would enable an unprecedented ability to augment of the physical world and could revolutionize the way humans interact with their surrounding environment. Such applications would be extremely interesting in urban environments, especially as more integrated technologies (such as Google Glass) become widely adopted. To enable researchers to utilize such a system we have developed an open-source library, Theia[2], that provides SLAM and SfM algorithms with a simple interface. We plan to release this open-source library this fall, with all future implementations related to this dissertation incorporated into the library.

We will also provide a mobile prototype of our system on an Android tablet that communicates with a server that we will maintain for public use. No special hardware is required for our system, so we will offer this prototype as a public app for general use.

It will be very important to determine how users interact with such systems. We will use our prototype to perform several key user studies to observe human interaction with our system. Tuite et al. [18] have demonstrated that users are willing to contribute to such a model if given incentives (they provided a game with more points given to "better" images). However, their game used static images and a slow offline process to add data to the global point cloud. An evaluation of a real-time system has yet to be considered. These studies will be crucial in determining design requirements for AR systems immersed in urban environments.

[2] http://jventura.mat.ucsb.edu/~cmsweeney/theia

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Commun. ACM*, 54 (10):105–112, Oct. 2011.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. IEEE, 2012.

[3] G. Bleser, H. Wuest, and D. Strieker. Online camera pose estimation in partially known and dynamic scenes. In *Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on*, pages 56–65. IEEE, 2006.

[4] V. Fragoso and M. Turk. Swigs: A swift guided sampling method. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, june 2013.

[5] S. Gauglitz, C. Sweeney, J. Ventura, M. Turk, and T. Hollerer. Live tracking and mapping from both general and rotation-only camera motion. *International Symposium for Mixed and Augmented Reality*, 2012.

[6] Y. Genc, S. Riedel, F. Souvannavong, C. Akinlar, and N. Navab. Marker-less tracking for ar: A learning-based approach. In *Mixed and Augmented Reality, 2002. ISMAR 2002. Proceedings. International Symposium on*, pages 295–304. IEEE, 2002.

[7] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2599 –2606, june 2009.

[8] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225 –234, nov. 2007.

[9] M. Klopschitz, A. Irschara, G. Reitmayr, and D. Schmalstieg. Robust incremental structure from motion. In *Proc. 3DPVT*, volume 2, 2010.

[10] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[11] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Proceedings of the 11th European conference on Computer vision: Part II*, ECCV'10, pages 791–804, Berlin, Heidelberg, 2010. Springer-Verlag.

[12] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV'12: Proceedings of the 12th European conference on Computer vision: Part II*, Oct. 2012.

[13] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1043–1050. IEEE, 2012.

[14] J. Park, S. You, and U. Neumann. Natural feature tracking for extendible robust augmented realities. In *Proc. Int. Workshop on Augmented Reality*, 1998.

[15] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 667 –674, nov. 2011.

[16] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *Proceedings of the 11th European conference on Trends and Topics in Computer Vision-Volume Part II*, pages 267–281. Springer-Verlag, 2010.

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.

[18] K. Tuite, N. Snavely, D.-y. Hsiao, N. Tabing, and Z. Popovic. Photocity: training experts at large-scale image acquisition through a competitive game. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1383–1392, New York, NY, USA, 2011. ACM.

[19] J. Ventura and T. Hollerer. Wide-area scene mapping for mobile visual tracking. *International Symposium for Mixed and Augmented Reality*, 2012.

[20] C. Wu. Towards linear-time incremental structure from motion. In *3DV, 2013.*, 2013.

[21] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 255–268, Berlin, Heidelberg, 2010. Springer-Verlag.