Visionary Collaborative Outdoor Reconstruction using SLAM and SfM

Philipp Fleck, Dieter Schmalstieg and Clemens Arth *
Graz University of Technology

ABSTRACT

In this position paper, we argue about a concept for collaborative outdoor reconstruction using SLAM clients and a scalable SfM engine running in the cloud. Based on previous observations and results, we discuss issues like illumination changes, overall scalability or the dacay of buildings, having a serious impact on the practical feasibility of such a system. Revisiting ideas and insights from work on outdoor reconstruction and localization done in the last couple of years, we outline an idea for collaborative and vivid reconstruction of the world, potentially through the cameras of millions of mobile devices.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented and virtual realities; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking; J.7 [Computer Applications]: Computers in Other Systems—Real time

1 INTRODUCTION

Today, we are surrounded by a "swarm" of smart cameras formed by the incredible number of mobile devices. Considering these smart cameras to observe the world almost constantly while in use, we can think about collaborative *Simultaneous Localization And Mapping* (SLAM) to capture a representation of the world, going far beyond what any single mobile device could do by itself. With tracking and mapping on such a massive scale, applications for Augmented Reality (AR) and 3D navigation would be raised to a completely new level of quality in user experience, accuracy and overall usefulness.

From recent literature in SLAM and *Structure from Motion* (SfM), realizing such a vision seems mainly a matter of effort, certainly requiring a horde of developers and substantial funds. The basic algorithms were proposed over the last couple of years (see *e.g.* [1, 3, 4, 6, 8, 14]), and the ever increasing computational power of the cloud and constantly improving wireless network quality further strengthens the basic feasibility of such a concept. So why is it still a vision and not a reality?

There is a multitude of hidden issues. One of them is scalability, another one is the changing appearance and shape of our surrounding, the illumination conditions, the weather and all related issues arising from these dynamics. Through the rest of this paper, we are going to describe our visionary concept, based on recent work on collaborative SfM and SLAM, and insights from the past years on outdoor reconstruction and localization in AR. Admittedly, we don't have a solution to all problems, however, we believe that the remaining issues might become minor once the major ones are adequately solved.

2 COLLABORATIVE SLAM AND SFM

In the past we worked on a system which is shortly described in the following (see Fig. 1 and [5] for more details). It serves as the starting point for the discussion of our novel concept.

*e-mail: {philipp.fleck,schmalstieg,arth}@icg.tugraz.at

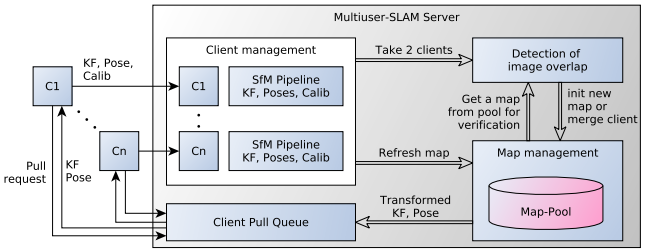

Figure 1: Client-server system architecture. $C_1 \ldots C_n$ represent SLAM clients, submitting information (*e.g.* keyframes) to the SfM server. The server creates a *client management* module for each client, where a new map is reconstructed and stored in the *map pool*. The server continuously updates the client maps and attempts to merge multiple maps, pushing relevant information (*e.g.* keyframes with camera poses) into the *client pull queues*, from where they can be pulled by clients to update their own local map.

Given a server running a SfM pipeline and multiple clients running SLAM, the reconstructions created by clients and server (i) use different feature descriptions, (ii) reside in different coordinate systems, and (iii) are created asynchronously, using per-node strategies involving global or local optimization. Clients and the server communicate using a protocol focusing on keyframes and camera poses only. Clients register their ID and provide their internal camera calibration parameters. The server initializes a per-client message queue and after initializing its local SLAM map, the client submits the corresponding stereo keyframe pair to the server, which reconstructs a per-client map independently of the client's own map.

During operation, the client asynchronously pulls messages from its queue (*e.g.* during any idle times while standing still). Upon certain events, the server puts relevant information into the client's queue. For example, if a second client transmits a keyframe that allows the merging of two clients' maps, the server offers additional keyframes and corresponding poses. The client may incorporate these keyframes into its local map. Additionally, the server may provide anchor points that allow for synchronizing the reference coordinate systems between multiple client instances.

Our system enforces *eventual consistency* between the server and its clients over time. We synchronize corresponding client and server maps by applying the *anchor point* method of Ventura *et al.* [14]. For each remote client map, the server determines a set of well-converged 3D map points that can be used to align the corresponding local client map. These anchor points can be integrated into the local client maps as fixed points and provide strong constraints in the clients' bundle adjustment optimization. In particular, a consistent reference coordinate system is established, which is essential for collaborative AR applications, where multiple clients render virtual objects in a consistent way.

3 A FLOATING 3D PICTURE OF THE WORLD

With such a system, multiple clients can collaboratively map and track within basically infintely large environments (see Fig. 2). Throughout the following paragraphs, we discuss several considerations for practical realization, going beyond known concepts in [10, 11, 16].

The Cloud A scalable SfM setup, as discussed above, can be implemented in the cloud. From the contributions of individual

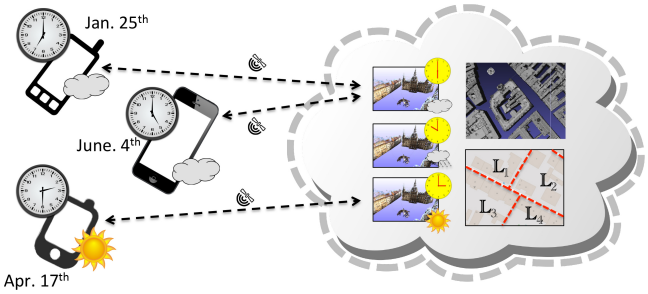

Figure 2: Conceptual system setup. While the cloud holds multiple reconstructions of the world for different weather and daytime conditions, clients exchange information based on their position and similar environmental conditions. All reconstructions are globally aligned to building models from OpenStreetMap. The 'world' itself is partitioned into multiple, potentially overlapping areas based on visibility considerations.

clients, reconstructions are generated and initially aligned along the GPS coordinates provided. While this adds an initial scale estimate to any of the local reconstructions on the clients, additional environmental information, such as, for example, building models available from OpenStreetMap or other cadastral maps, are used to align the reconstructions in an accurate manner [7]. Multiple reconstructions are kept for different weather and daytime conditions, while clients can leverage the already available information, encountering similar environmental conditions.

Partitioning the overall model of the world into individual blocks based on visibility information [2] further supports scalability of the setup. This concept has proved to be advantagous in MMORPGs (*Massively Multiplayer Online Role Playing Games*) like *World of Warcraft*[1]. GPS serves as a real-time partition selector. From previous studies on the availablity of data in public spaces [1] certain areas are covered by a large number of overlapping observations, for example, in sightseeing places. However, other areas are only sparsely captured, if at all. Depending on the availability of data, reconstructions are newly created or updated using change detection algorithms [12, 13], to tackle changes caused by the decay of buildings or illumination.

The Clients Nowadays mobile devices are equipped with sufficient computational and memory resources to generate SLAM maps autonomously. Additional information is retrieved on demand from the cloud to further improve the local 3D model of the environment and to align it globally. Through this local alignment, inconsistencies are detected on the client - or merely indicated - to signalize changes in the environment back to the cloud for it's further reduction. Thereby, not only changes to static scenery are detected, but also enhanced Computer Vision technology is used to detect moving pedestrians or cars [15], to further improve the reconstruction or create an analog virtual representation of the world's dynamics in the cloud [9].

Always-on Devices While the concept outlined as such refers to mobile smartphones mainly, it arguably becomes more plausible considering always-on devices, such as head-worn displays (HMDs) or data glasses (*e.g.* Google Glass). People are occasionally using their mobile devices, but not primarily for any application related to SLAM technology. With the increasing interest in Virtual Reality (VR), head-worn devices might become mainstream in the next decade. As a constant stream of image and sensor information is generated, using SLAM becomes natural.

In the meantime, we envision services to take pictures with the back camera of mobile devices while making calls. As an analogy to home screen advertising services[2], these images are sent to the cloud, to at least serve the cloud with actual data to further pursue the creation of a floating 3D picture of the world. Naturally, both approaches, local SLAM and occasional image transmission, seamlessly integrate into the same framework.

Privacy and Legal Concerns Legal regulations are recently pushed forward by innovations with overwhelming mass adoption. Changes in driving regulations triggered by the advent of Google Glass[3], imply further changes, as we increasingly create data and observe the world digitally or become digitally observed. We have to face the fact that we are responsible for our own privacy, *i.e.* either use such a service or not. We argue that, overall, this remains a minor issues to be solved once we create a large-scale implementation of the concept.

4 Concluding Remarks

The concept discussed within this paper considers a large number of algorithms and approaches. We argue that the implementation of such a system is definitely feasible, once a concrete business case is developed. To us the advent of such a system remains mainly a matter of time, as new innovations in usage and convenience of mobile gadgets revolutionize our daily life in the near future.

Acknowledgements

This work was partially funded by the Christian Doppler Laboratory on Semantic 3D Vision.

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] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide Area Loc. on Mob. Phones. In *ISMAR*, pages 73–82, 2009.

[3] A. J. Davison. Real-Time Simultaneous Localisation and Mapping with a Single Camera. In *ICCV*, pages 1403–1410, 2003.

[4] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *ECCV*, September 2014.

[5] P. Fleck, C. Arth, C. Pirchheim, and D. Schmalstieg. Track. and mapping with a swarm of het. clients. In *ISMAR*, pages 136–139, 2015.

[6] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *ISMAR*, Nara, Japan, November 2007.

[7] P. Lothe, S. Bourgeois, F. Dekeyser, E. Royer, and M. Dhome. Towards geographical referencing of monocular SLAM reconstruction using 3d city models: Application to real-time accurate vision-based localization. In *CVPR*, pages 2882–2889, 2009.

[8] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3d reconstruction. In *CVPR*, CVPR '06, pages 363–370, Washington, DC, USA, 2006.

[9] U. Neumann, S. You, J. Hu, B. Jiang, and I. O. Sebe. Visualizing reality in an augmented virtual environment. *Presence: Teleoper. Virtual Environ.*, 13(2):222–233, Apr. 2004.

[10] L. Riazuelo, J. Civera, and J. Montiel. C2TAM: A Cloud framework for cooperative tracking and mapping. In *Robotics and Autonomous Systems, Volume 62, Issue 4, Pages 401-413, April 2014*, 2014.

[11] C. Sweeney. Improved Outdoor Augmented Reality through Globalization. In *Doctoral Consortium, ISMAR*, 2013.

[12] A. Taneja, L. Ballan, and M. Pollefeys. Image based det. of geometric changes in urban env. In *ICCV*, pages 2336–2343, Nov 2011.

[13] A. O. Ulusoy and J. L. Mundy. Image-based 4-d reconstruction using 3-d change detection. In *ECCV*, pages 31–45, 2014.

[14] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. Global Local. from Mon. SLAM on a Mob. Phone. *TVCG*, 20(4):531–539, 2014.

[15] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.

[16] D. Zou and P. Tan. CoSLAM: Collaborative Visual SLAM in Dynamic Environments. *PAMI*, 35(2):354–366, 2013.

[1] http://eu.battle.net/wow

[2] http://www.infoworld.com/article/2608047/mobile-apps/there-s-no-escape--ads-come-to-your-smartphone-screen.html

[3] http://bgr.com/2014/02/25/google-glass-driving-laws/