Towards Dense Collaborative Mapping using RGBD Sensors

Louis Gallagher*†

and John B. McDonald

Department of Computer Science, Maynooth University, Co. Kildare, Ireland.

Abstract

Development of collaborative, perception driven autonomous systems requires the ability for collaborators to compute a rich, shared representation of the environment, and their place in it, in real-time. Using this shared representation, collaborators can communicate geometric, semantic and dynamic information about the environment across frames of reference to one another. Existing state-of-the art dense mapping systems provide a good starting point for developing a collaborative mapping system, however, no system currently covers collaborative mapping directly. In this paper, we introduce our approach to dense collaborative mapping, offering an introduction to the problem, a discussion of the key challenges involved in developing such a system and an analysis of preliminary results.

Keywords: Dense, SLAM, Reconstruction, Mapping, Collaborative.

1 Introduction

The aim of dense visual simultaneous localisation and mapping (VSLAM) is to recover a dense reconstruction of a scene from a freely moving visual sensor. This is achieved through the continued fusion of measurements into a single representation, whilst simultaneously tracking the motion of the sensor, all in real-time. The state-of-the-art in the field includes a multitude of systems offering large-scale, high-precision mapping and tracking capabilities [Dai et al., 2017, Whelan et al., 2015, Kerl et al., 2013, Whelan et al., 2014, Newcombe et al., 2011, Engel et al., 2014]. To date all of these dense SLAM systems have focused on single sensor mapping. In this paper we report on initial work to address the wider problem of collaborative, multi-sensor mapping and tracking which is essential to many robotics and augmented reality applications such as human robot interaction (HRI), cooperative robotics and multi-session mapping. Historically, collaborative mapping has been a recurring theme in the SLAM literature [Saeedi et al., 2016], though the concept has yet to be extended to the more contemporary setting of dense visual SLAM.

This paper reports on first results of in-progress research to extend the Elastic Fusion (EF) single sensor dense mapping system to allow for multi-sensor collaborative mapping and tracking using RGBD sensors. The contributions of the paper are: (i) a discussion of the challenges involved in extending EF to allow for full multi-sensor collaborative mapping, (ii) a description of our proposed multi-sensor EF framework and the components implemented to date; and (iii) a qualitative comparison between multi and single sensor EF.

The remainder of the paper is structured as follows. Section 2 provides a brief summary of the elements of the EF algorithm pertinent to understanding our proposed extensions. Section 3 discusses the challenges and proposed solutions to allow this extension, and provides details of the subset of the solutions that we have implemented to date. Section 4 presents the multi-sensor dense mapping capabilities of the current system, and provdes an initial comparison of its outputs to single-sensor EF. Finally, in Section 5 we give concluding remarks and discuss future research directions.

*This research is funded by the Irish Research Council, Project ID: GOIPG/2016/1320

†louis.gallagher.2013@mumail.ie

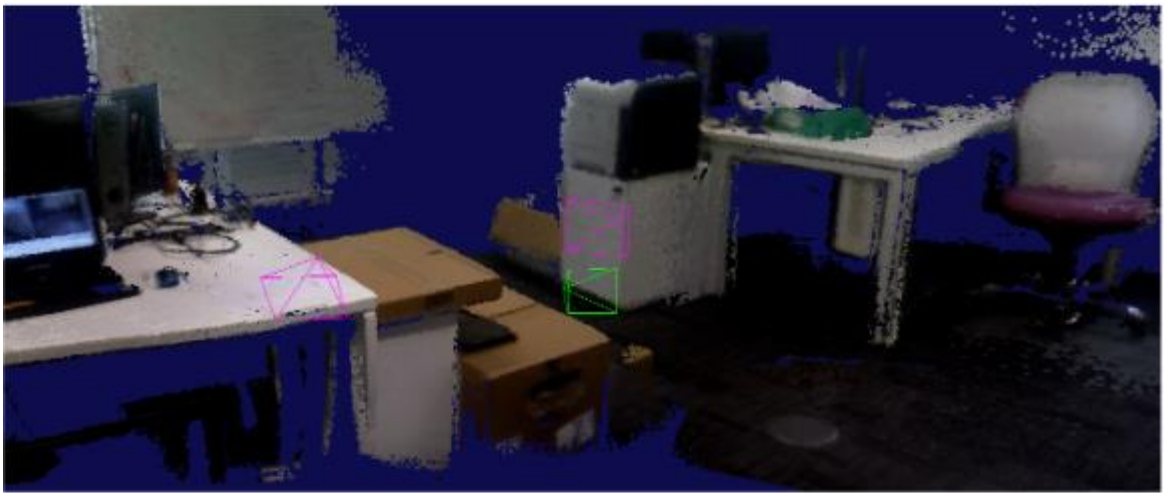


Figure 1: A three sensor collaborative mapping session in our system. All sensors started with the same initial pose but proceeded along independent trajectories

2 Background

In this section we give an overview of the EF dense mapping system, although the reader will find a more comprehensive treatment in [Whelan et al., 2015]. EF takes a point-based fusion approach to dense mapping. As an RGBD sensor traverses an environment its measurements are fused into a single model, internally represented by an unstructured list of surface elements (surfels),M. Each surfel in M is an estimate of a discrete point on the underlying continuous surface contained in the environment being mapped. M is split along temporal lines into two distinct sublists; Θ, containing active surfels that have been observed recently, and Φ, containing inactive surfels that have not been observed in a period of time δ_t.

At each time step, the global pose of the sensor, $P_t \in SE_3$, at time t is estimated by aligning the active model-predicted surface, derived from Θ using P_{t-1}, to the surface contained in the latest sensor frame at time t. This alignment yields a transformation that is applied to P_{t-1} to give P_t. Once $P_t$is resolved the latest frame can be fused into Θ, and thus into M.

Assuming that a global loop closure has not occured at the current time step, a local loop closure between Θ and Φ is sought, reactivating inactive surfels and maintaining local surface consistency. Local loops are closed by performing a surface registration between the portion of Θ in view of P_t to the portion of Φ in the same view.

Occassionally, the sensor drifts too far for the estimated model to be corrected by the local loop closure mechanism. To solve this problem a randomised fern-encoding database containing discriminative views of the scene is maintained. The database is searched for a view matching the current active model-predicted view. If a matching view is found then a global loop closure is determined by solving a model-to-model surface registration between the surfaces underlaid within the matching views, akin to local loop closure.

In the case that either a local or a global loop closure is achieved the model is non-rigidly deformed by applying a space deforming graph to M [Whelan et al., 2015].

Surface registrations in EF are performed using a local alignment technique. An objective function is defined over both geometric and photometric constraints between the two surfaces, parameterised by T_k, the k_{th} estimate of the transformation $T \in SE_3$ that aligns them. This objective defines an error surface, E, with respect to T. Through iterative non-linear least-squares, using Gauss-Newton optimisation and a three-level coarse-to-fine pyramid scheme, the objective is minimized by updating T_k in a direction of descent along E, yielding increasingly refined approximations of

3 Extending ElasticFusion for Multi-Sensor SLAM

In extending EF to permit collaborative mapping we identify two distinct phases of processing for any given input stream. This distinction arises from the fact that in collaborative mapping there is, in general, no common frame of reference to begin with, and hence each sensor's initial pose is unknown relative to the other sensors. Therefore, at the outset, the system assumes that each sensor is positioned at the origin of a frame of reference that is independent to that of other sensors. During this initial phase each sensor input is essentially processed via an independent EF mapping pipeline. As mapping progresses the aim is for global inter-sensor loop closures to occur, thereby providing the necessary transformations between the sensor submaps. These transformations permit alignment of submaps into a common frame of reference which makes it possible to perform multi-sensor fusion into a single global map. In order to concentrate on the development required for this second phase of processing, in this paper we constrain our multi-sensor datasets such that each sensor starts with the same initial pose. Thus from the outset we assume a single global map with multiple independently moving sensors. Therefore, our system to date, deals with the post-alignment stage of collaborative mapping.

Hence, we take a phased approach to extending EF, where in the first phase of the extension we assume datasets that are constrained in the manner described above. Multiple sensors can then fuse measurements into, and track against, a common global surfel map. Our representation for multiple sensor mapping is a tuple of the form $<M,\{P_{it}\}>$ for surfel map M and sensor poses$\{P_{it}\}$, where each$P_{it}\in SE_3$represents the pose of sensor I at time t. Local surface consistency is maintained in the same manner as discussed in Section 2 with the exception that the active region of M contains the surfels in M that are active with respect to at least one sensor. For detecting global loop closures all sensors keep a common fern encoding database, loops can then be detected and closed in the same way as was described in Section 2.

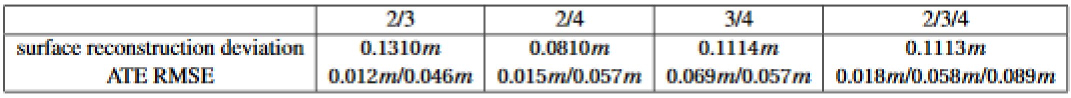
Following this, we will concentrate on seperating out the sensor's temporal windows such that each sensor defines its own active and inactive map regions. Having one active region for all sensors leads to inefficient view predictions and temporal incoherence, preventing sensors from closing local loops in certain situations. For example when two sensors following independent trajectories intersect, with each sensor transitioning to a portion of the map that the other has kept active, no local loop closures will occur. Once the temporal windows have been separated multi-sensor mapping and tracking can then continue as before. To maintain separate temporal windows for each sensor we introduce the concept of a context. A context is a triple of the form $<\Theta_{it}, \Phi_{it}, P_{it}>$, where $\Theta_{it} \subset M$ and $\Phi_{it} \subset M$ represent the regions of M that are active and inactive w.r.t sensor I at time t. As before $P_{it} \in SE_3$ represents the pose of sensor I at time t. Thus collaborative mapping can be represented as a tuple of the form $<M, \{\zeta_i\}>$, for surfel map M and contexts $\{\zeta_i\}$.

In the final phase of extension we will focus on removing the constraint that each sensor must start with the same initial pose. Thus, initially each sensor will be associated with its own frame of reference denoted by $F_i = <M_i, \{\zeta_{ij}\}>$ for surfel map M_i and contexts $\{\zeta_{ij}\}$. By solving the global alignment problem between maps, frames of reference can be merged, and so, as mapping proceedes and inter-map loop closures occur, our representation will tend towards a single global map.

4 Experiments

In this section we report the first results of our approach. In the absence of ground-truth multi-sensor VSLAM datasets we use single-sensor EF to generate scene reconstructions and sensor poses. We then measure the deviation between these reconstructions and pose estimates and those outputted by our system as we incrementally increase the number of sensors collaboratively mapping. All experiments were run on a machine with an NVidia GeForce GTX 1080 ti GPU, 11GB of GDDR5 VRAM, an 8 core Intel i7-7700K CPU running at4.20GHz and 16GB of DDR4 system memory.

We give a qualitative comparison between our system and EF using a custom lab dataset. The dataset consists of four sequences through the same scene. Each sequence was captured with an ASUS Xtion pro live depth sensor running at30hz. The initial pose of the sensor is the same across all sequences. The first sequence is used in conjunction with single-sensor EF to compute a reconstruction of the scene. We also use single-sensor EF to compute sensor poses for the other three sequences. Then, we compare the single-sensor reconstruction and sensor poses to those computed by our system under the different permutations of the other three sequences collaboratively mapping and tracking. Table 1 summarises this data. The reader is encouraged to watch the accompanying video for a clearer visualisation of both this dataset and dense collaborative mapping in general (https://youtu.be/qYNpP 5Vp7I).



	2/3	2/4	3/4	2/3/4
surface reconstruction deviation	0.1310m	0.0810m	0.1114m	0.1113m
ATE RMSE	0.012m/0.046m	0.015m/0.057m	0.069m/0.057m	0.018m/0.058m/0.089m

Table 1: The first row gives the numbers of the sequences being used in the collaboration. The second row gives the per collaboration mean distance from each point to the nearest point in the single-sensor reconstruction. In the last row we give the per sequence ATE RMSE for each collaboration.

5 Discussion

We have reported on our initial work on a dense collaborative mapping system, demonstrating its capabilities through a comparative analysis with EF. In future work we will focus on improving the temporal coherence of our system and increasing its generality. To address the former issue, each sensor will define its own active map region, allowing us to maintain local surface consistency even in cases where sensors enter each others active regions. To address the latter issue we plan to leverage a global alignment technique, fast global registration [Zhou et al., 2016], to align sensor specific maps into a single global map on the fly. An important aspect of future work will be the creation of synthetic ground-truth datasets, allowing us to measure the performance of our system.

References

[Dai et al., 2017] Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., and Theobalt, C. (2017). Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*,36(3):24:1–24:18.

[Engel et al., 2014] Engel, J., Schöps, T., and Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision (ECCV).*

[Kerl et al., 2013] Kerl, C., Sturm, J., and Cremers, D. (2013). Dense visual slam for rgb-d cameras. In *Proc.of the Int. Conf. on Intelligent Robot Systems (IROS).*

[Newcombe et al., 2011] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J.,Kohli, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE Intl. Symposium on Mixed and Augmented Reality*, ISMAR '11, pages 127–136, Washington, DC, USA. IEEE Computer Society.

[Saeedi et al., 2016] Saeedi, S., Trentini, M., Seto, M., and Li, H. (2016). Multiple-robot simultaneous local-ization and mapping: A review. *J. of Field Robotics*, 33(1):3–46.

[Whelan et al., 2014] Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J., and McDonald, J.(2014). Real-time large scale dense RGB-D SLAM with volumetric fusion. *Intl. J. of Robotics Research, IJRR.*

[Whelan et al., 2015] Whelan, T., Leutenegger, S., Moreno, R. S., Glocker, B., and Davison, A. (2015). Elasticfusion: Dense slam without a pose graph. In *Proceedings of Robotics: Science and Systems*, Rome, Italy.

[Zhou et al., 2016] Zhou, Q., Park, J., and Koltun, V. (2016). Fast global registration. *In European Conference on Computer Vision (ECCV),* pages 766–782.