TITLE

Anonymous ICCV submission

Paper ID ****

Abstract

The availability of high quality, consumer grade, depth sensing equipment has motivated much research in dense SLAM systems for object reconstruction. Though much progress has been made, the well known loop closure problem is still an active area of research. In this paper we present an object reconstruction system that is capable of counteracting camera tracking drift and detecting loop closures, such that high quality reconstructions can be yielded even with a high level of interaction. We demonstrate that our novel correction algorithm is able to reduce tracking drift, detect loop closures and increase overall object reconstruction quality.

0.1. Introduction

Dense SLAM(Simultaneous Localisation and Mapping) has proven to be an effective paradigm for the reconstruction of scenes of moderate size, with much research on the topic driven by the availablility of consumer grade depth sensing equipment. However, there is a heavy reliance on descriptive geometry in the scene when there is a lack of texture. Less descriptive geometry leads to an increase in camera tracking error and causes inconsistencies when a loop closure event occurs.

As object reconstruction can be seen as a smaller scale equivelant of the scene based dense reconstruction problem, it too is prone to the tracking drift and loop closure problem, sometimes to a prohibitive level. Often it may be desirable to perform object reconstruction in an interactive way, for example, as a component of a scene understanding system, or to procure training data for the object in question. With a high level of interaction comes an exacerbation of the aformentioned shortcomings of dense SLAM, particulary due to the potential for frequent, repetetive motion. This is the problem that is addressed in this work.

In this paper we present a probabilistic object reconstruction framework for in hand reconstruction of rigid objects based on object appearances. The framework facilitates the correction of camera tracking drift by representing the object to be reconstructed as a collection of overlapping subsegments such that deformations may be inferred to keep the subsegments aligned, resulting in a consistent overall model. We utilise a volumetric representation for each of these object subsegments, as with many larger scale reconstruction systems. Each voxel in the subsegments has additional appearance posterior information pertaining to the voxels membership of the object. Over time, multiple volumes containing both surface and probabilistic appearance information are maintained and manipulated to yield a robust and temporally consistent model.

In addition, to further increase the robustness of the object tracking, a novel tracking procedure is used, utilising appearance features in image space. Finally, the optimum object shape is optimised for within a continuous max flow framework.

1. Background and Pertinent Literature

1.1. Dense 3D Reconstruction

Much 3D reconstruction work in recent years has been influenced by the seminal KinectFusion[4] of Newcombe et al in which RGBD data was integrated into a volumetric reporesentation of the scene, performing simultaneous tracking and mapping. The end result was high quality 3D static scene models. However KinectFusion notably suffered from tracking drift and had no capacity to handle loop closure events.

Following KinectFusion, Neissner et al present another volumetric reconstruction system[5] based around the notion of hashing regions of space in to voxel blocks. The primary contribution is the ability to scale the abilities of KinectFusion to larger scenes. However, the contributions do not extend to the camera tracking drift and loop closure problems.

Prisacariu et al present an alternative voxel hashing based system[6] which provides many optimisations and an open source implementation. However, the limitations of[4, 5] with regards to loop closure and tracking drift are still present. However, a later publication[2] from the authors of[6] presents a loop closure and tracking drift reduc-

tion solution based on the splitting of the scene in to sub scenes with pose adjustments made to tracking constraints between them. This approach is pertinent to this work as it inspired the multi subsection approach that we take.

1.2. Object Reconstruction

In addition to the larger scale dense SLAM works discussed in the previous section, there has been much work on object reconstruction and object centric SLAM. Ren et al present a probabilistic object tracking and reconstruction system[9] that like our work builds object reconstructions based on an appearance model. The presented system of Ren et al evolves a level set object representation for voxels that are on the object, as per the appearance model. However, the presented system does not have any provision for loop closure detection and is prone to tracking drift over time.

Kolev et al present a probabilistic 3D segmentation and surface extraction algorithm[3] based on a variational evolution of a level set representation. Object appearance probabilities are fused in to the objects volume for the segmentation, much like the approach taken in this paper, such that the algorithm is robust to outliers in the observation images. However, their system does not provide any handling of loop closure occurrences and the paper makes no reference to tracking integrity. In addition, the images on which their algorithm was tested contained only the objects to be reconstructed, with no background noise.

Another volumetric object reconstruction system is presented by Gupta et al[1], using monocular multi view cues. The authors perform object segmentation within a graph cut framework to yield object models and perform tracking based on visual and textual cues. However the authors report fluctuating camera tracking quality due to the breakage of brightness constancy and specular surfaces. In addition, as with all other works introduced up to this point there is no loop closure ability and tracking drift is an issue.

Slavcheva et al present a volumetric object reconstruction system for RGBD sensors in which pose estimates are yielded by registering pairs of SDF(Signed Distance Function) volumes. Unlike many of the aforementioned works, the authors do present a loop closure step, however it is performed offline as a post processing step, as such it is unclear what the timings for this step are.

Finally, Weise et al present an explicit, surfel based object reconstruction system[8] based on objects rotating in front of a 3D range scanner. During reconstruction an object topology graph is constructed that is used on-line to handle loop closure cases. When a loop closure is detected, if there are discrepancies in the topology graph then deformations are applied locally to patches of the object surface. As such, the two misaligned ends of the surface are realigned. However, what is not clear is the systems ability to handle motion beyond a simple rotation. In addition, the use of the explicit Surfel based representation makes the reconstructed models less amenable to structured computation(such as processing by a Convolution Neural Network) than their volumetric counterparts.

2. Algorithmic Details

2.1. Representation and Fusion Procedure

The proposed system is inspired by[3] in that the representation used for the shape of the object to be modelled is a volume of probabilities, pertaining to posteriors over a voxels assignment to being either belonging to the object or not. In the proposed system this volume of posterior probabilities is built into with each frame, parallel to the fusion process in systems such as KinectFusion[4] and InfiniTAM[6].

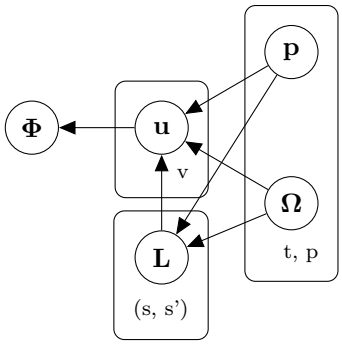
The probabilities that are accumulated into the volume are generated from a Random Forest based appearance model using patch based features encompassing appearance and surface information, such as depth gradients, initialised prior to reconstruction by a user interaction in the first frame. There are two classes in the appearance model, one for the foreground object and one for the background, with the foreground object indicated by a bounding box on the first RGB frame. At each frame a smaller, image sized volume is constructed based on the predictions of the current frame. During the fusion process, this smaller volume is mapped in to as a source of voxel wise appearance probability information. The overall appearance based posterior for a given voxel $\psi \in \mathbf{\Psi}$ takes the following form, where $\mathbf{\Psi}$ is the volume of voxels for which measurements are accumulated, $\mathbf{\Phi}$ is the volume of voxels pertaining to the object, $\mathbf{\Omega}$ is the current image observation and $\mathbf{p}$ is the currently tracked pose:-

$$P(\psi \in \mathbf{\Phi}|\mathbf{\Omega}, \mathbf{p}) = \prod_{t=0}^{\infty} P(\psi_t \in \mathbf{\Phi}_t|\mathbf{\Omega}_t, \mathbf{p}_t) \qquad (1)$$

The above encodes the probability of a voxel belonging to an object as the product of the instantaneous conditionals for observations at each time step.

2.2. Probabilistic Formulation of Object Reconstruction

As previously highlighted, central to the proposed system is a volume of posterior probabilities pertaining to a voxels membership of an object point versus a background scene point. This allows one to formulate the full joint distribution over the object as the following Probabilistic Graphical Model:- Where $\mathbf{\Phi}$ is the shape to be reconstructed, $\mathbf{u}$ is the appearance model volume, $\mathbf{L}$ is the set of consistency constraints for each adjacent sub volume pair, $\mathbf{\Omega}$ is the set of RGBD image pixels and $\mathbf{p}$ the set of poses over time.



This gives rise to the following analytical formulation of the above distribution:-

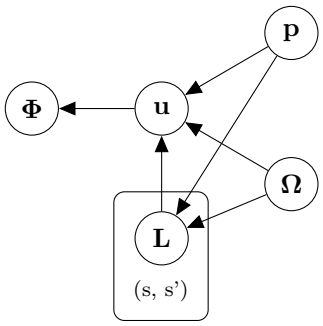
$$P(\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{p}, \mathbf{u}, \mathbf{L}) = \prod_{v \in \mathcal{V}} P(\mathbf{\Phi}|\mathbf{u}_v) \prod_{t=0}^{\infty} \prod_{p \in \mathcal{P}} \prod_{(s,s') \in \mathcal{S}} P(\mathbf{u_v}|\mathbf{\Omega}_{p,t}, \mathbf{p}_t, \mathbf{L}_{(s,s')}) P(\mathbf{L}_{(s,s')}|\mathbf{\Omega}_{p,t}, \mathbf{p}_t) P(\mathbf{L}_{(s,s')}) P(\mathbf{p}_t) P(\mathbf{\Omega}_{p,t}) \tag{2}$$

Where $\mathcal{V}$ is the set of voxels across all sub volumes, $\mathcal{P}$ is the set of RGBD pixels for a given frame and $\mathcal{S}$ is the set of sub volumes.

However, if one were to assume temporal and pixel wise independence in the RGBD observations and temporal independence in the poses, the plate containing $\mathbf{\Omega}$ and $\mathbf{p}$ can be removed:-

$$P(\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{p}, \mathbf{u}, \mathbf{L}) = \prod_{v \in \mathcal{V}} P(\mathbf{\Phi}|\mathbf{u}_v) \prod_{(s,s') \in \mathcal{S}} P(\mathbf{u_v}|\mathbf{\Omega}, \mathbf{p}, \mathbf{L}_{(s,s')}) P(\mathbf{L}_{(s,s')}|\mathbf{\Omega}, \mathbf{p}) P(\mathbf{L}_{(s,s')}) P(\mathbf{p}) P(\mathbf{\Omega}) \tag{3}$$

Furthermore, if one assumes voxel wise independence, the plate over voxels can be removed. Finally, assuming $P(\mathbf{p})$ and $P(\mathbf{\Omega})$ are uniform distributions, then we have the following, simpler distribution:-



With the following analytical form:-

$$P(\mathbf{\Phi}, \mathbf{\Omega}, \mathbf{p}, \mathbf{u}, \mathbf{L}) = P(\mathbf{\Phi}|\mathbf{u}) \prod_{(s,s') \in \mathcal{S}} P(\mathbf{u}|\mathbf{\Omega}, \mathbf{p}, \mathbf{L}_{(s,s')}) P(\mathbf{L}_{(s,s')}|\mathbf{\Omega}, \mathbf{p}) P(\mathbf{L}_{(s,s')}) \tag{4}$$

2.3. Inferring Deformations

The tracking consistency constraint denoted by the variable $\mathbf{L}$ in the above graphical model can be enforced in terms of minimising the transformations between adjacent submaps, such that the camera poses tracked in each subvolume are consistent. This follows the approach of[2]. However, the approach proposed in this work differs in that the optimisation is integrated in to the probabilistic formulation previously outlined. Given instantaneously inferred transforms between subvolumes obtained from tracking results, the objective is to infer a robust, consistent deformation transformation for the subvolume pair.

As such, for each pair of visible subvolumes (s, s'), the following posterior must be maximised:-

$$P(\mathbf{\Omega}, \mathbf{p}|\mathbf{L}_{(s,s')}) = \frac{P(\mathbf{L}_{(s,s')}|\mathbf{\Omega}, \mathbf{p}) P(\mathbf{\Omega}|\mathbf{p}) P(\mathbf{p})}{P(\mathbf{L}_{(s,s')})} \tag{5}$$

The intuition behind the above equation is that the deformation $\mathbf{L}_{(s,s')}$ applied to the probability field $\mathbf{u}$ should increase the probability of observing the current pose $\mathbf{p}$ given the current RGBD frame $\mathbf{\Omega}$ by reducing the variance of the camera tracking result. As such, global tracking variance is reduced by enforcing local consistency, improving the quality of the reconstruction.

The maximisation of the above posterior consists of a two stage process. Firstly, as a preprocessing step, the true transformation between subsegments is estimated over time from observed offsets, as documented in the following section. This inferred transformation is used to initialise a gradient based maximisation of the above posterior to yield an optimal deformation. The intuition behind this schema is that the gradient based optimisation routine is used to refine the inferred transformations between subsegments, improving robustness.

The following proportionality to the distribution over deformations is made:-

$$P(\mathbf{L}_{(s,s')}|\mathbf{\Omega}, \mathbf{p}) \propto P(F_s(\mathbf{x})|F_{s'}(G(\mathbf{x}))) \tag{6}$$

With the likelihood function taking the following form(a Gaussian distribution is assumed):-

$$P(F_{s'}(G(\mathbf{x}))) = \prod_{(s,s') \in \mathcal{S}} \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-(F_s(\mathbf{x}) - F_{s'}(G(\mathbf{x})))^2}{2\sigma^2} \tag{7}$$

Or alternatively:-

$$\ln P(F_{s'}(G(\mathbf{x}))) = m \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{(s,s')\in\mathcal{S}} \left(F_s(\mathbf{x}) - F_{s'}(G(\mathbf{x})) \right)^2 \quad (8)$$

Where $F(.)$ is a scalar valued SDF(Signed Distance Function), $\mathbf{x}$ is a point represented by a 3-vector, and $G(.)$ is a transformation function taking the following form:-

$$G(\mathbf{x}) = \mathbf{R}(r_1, r_2, r_3)\mathbf{x} + \mathbf{t} \quad (9)$$

Where $\mathbf{R}(.)$ is a rotation matrix from the Special Orthogonal group $\mathbb{SO}(3)$ paramaterised by the three Modified Rodrigues Parameters[7] r_1, r_2 and r_3.

Note that the logarithmic form of the above likelihood is suitable to Nonlinear Least Squares optimisation, allowing the posterior of equation 4 to be maximised in terms of the likelihood term of equation 4. To perform MLE(Maximum Likelihood Estimation) over this likelihood using an optimisation routine such as Levenberg Marquardt, the following gradients must be computed for the rotational component of the deformation:-

$$\frac{\partial E}{\partial r_n} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial G}\frac{\partial G}{\partial r_n} \text{for } n \in 1, 2, 3 \quad (10)$$

Similarly for the translational component:-

$$\frac{\partial E}{\partial \mathbf{t}_d} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial G}\frac{\partial G}{\partial \mathbf{t}_d} \text{for } d \in x, y, z \quad (11)$$

2.4. Estimation of Inter-Subvolume Constraints

Due to the on-line nature of the fusion pipeline there is a degree of uncertainty around the measured offsets between subvolumes. As such, to counter this and yield a robust starting constraint for the previously described optimisation procedure, a recursive bayesian estimation procedure is used to estimate the true pose difference between two subvolumes.

As such, the update for the starting deformation transformation $\mathbf{L}_t$ may be defined in terms of $\mathbf{L}_{t-1}$ and observed offsets transforms between segments $\mathbf{Z}_{(\mathbf{s},\mathbf{s}')} = \mathbf{T}_s^{-1}\mathbf{T}_{s'}$ in a predict-update fashion, using the following prediction equation:-

$$P(\mathbf{L}_i|\mathbf{Z}_{1:i-1}) = \int P(\mathbf{L}_i|\mathbf{L}_{i-1})P(\mathbf{L}_{i-1}|\mathbf{Z}_{1:i-1})d\mathbf{L}_{i-1} \quad (12)$$

and update equation

$$P(\mathbf{L}_i|\mathbf{Z}_{1:i}) = \frac{P(\mathbf{Z}_i|\mathbf{L}_i)P(\mathbf{L}_i|\mathbf{Z}_{1:i-1})}{P(\mathbf{Z}_i|\mathbf{Z}_{1:i-1})} \quad (13)$$

where

$$P(\mathbf{Z}_i|\mathbf{Z}_{1:i-1}) = \int P(\mathbf{Z}_i|\mathbf{L}_i)P(\mathbf{L}_i|\mathbf{Z}_{1:i-1})d\mathbf{L}_i \quad (14)$$

However, it should be noted that the denominator of the update equation may in practice be replaced with a normalisation constant as follows:-

$$P(\mathbf{L}_i|\mathbf{Z}_{1:i}) = \alpha P(\mathbf{Z}_i|\mathbf{L}_i)P(\mathbf{L}_i|\mathbf{Z}_{1:i-1}) \quad (15)$$

Note that in the above equations, the constraint pair subscript has been dropped for clarity in notation.

This online prediction-update schema will allow the evolution of pose difference constraints over time increasing their robustness at the time of optimisation of deformations.

3. Volumetric Object Segmentation

The final stage in the proposed object reconstruction pipeline is the segmentation of the object voxels from those that have had measurements fused from the background. This process can be posed as an energy minimisation problem over a cut in voxel space, such that a segmentation in 3D is yielded. The following energy function consists of the unary posterior probabilities over appearance accumulated during the fusion process and an additional pairwise smoothing term representing the physical appearance similarity of the object region represented by the voxel:-

$$E() \quad (16)$$

4. Results

5. Discussion

References

[1] T. Gupta, D. Shin, N. Sivagnanadasan, and D. Hoiem. 3dfs: Deformable dense depth fusion and segmentation for object reconstruction from a handheld camera. *CoRR*, abs/1606.05002, 2016. 2

[2] O. Kähler, V. A. Prisacariu, and D. W. Murray. *Real-Time Large-Scale Dense 3D Reconstruction with Loop Closure*, pages 500–516. Springer International Publishing, Cham, 2016. 1, 3

[3] K. Kolev, T. Brox, and D. Cremers. Robust variational segmentation of 3D objects from multiple views. In K. F. et al., editor, *Pattern Recognition (Proc. DAGM)*, volume 4174 of *LNCS*, pages 688–697, Berlin, Germany, Sept. 2006. Springer. 2

[4] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 1, 2

[5] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *ACM Transactions on Graphics*, 32(6):169, 2013. 1

[6] V. A. Prisacariu, O. Kahler, M. M. Cheng, C. Y. Ren, J. Valentin, P. H. S. Torr, I. D. Reid, and D. W. Murray. A Framework for the Volumetric Integration of Depth Images. *ArXiv e-prints*, 2014. 1, 2

[7] M. D. Shuster. Survey of attitude representations. *Journal of the Astronautical Sciences*, 41:439–517, Oct. 1993. 4

[8] T. Weise, T. Wismer, B. Leibe, and L. V. Gool. In-hand scanning with online loop closure. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1630–1637, Sept 2009. 2

[9] C. Yuheng Ren, V. Prisacariu, D. Murray, and I. Reid. Star3d: Simultaneous tracking and reconstruction of 3d objects using rgb-d data. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2