Exploiting Depth Information from Tracked Feature Points in Dense Reconstruction for Monocular Cameras

Qirui Zhang[*1] Takafumi Taketomi[*1] Goshiro Yamamoto[*1]
Christian Sandor[*1] Hirokazu Kato[*1]

Abstract – In dense 3D reconstruction work for monocular simultaneous localization and mapping (SLAM), to present coherence and prevent abrupt change in reconstructed surfaces, we normally model the contextual constraint of physical properties in a neighbourhood of space as a certain prior smoothness term concisely into the optimization process. In our work, we first had a careful discussion about the trade-off between precision and accuracy for different prior smoothness terms and how these affected the optimization process of the depth map based on photo consistency measurement. We then presented a method which uses depth information of tracked feature points as priors in the optimization process. Finally, we verified effectiveness of our method by conducting quantitative evaluation experiments in a simulated environment. We also qualitative evaluation in a real environment. We confirmed that feature prior information can improve the accuracy of reconstructed structure at the strong texture area.

Keywords : structure from motion, 3D reconstruction, dense model, total variation, depth map, cost volume, monocular camera

1 Introduction

In computer vision, visual SLAM based on a monocular RGB camera, which means tracking a handheld camera and simultaneously recovering the three-dimensional structure of the environment in real-time is a challenging and promising direction and increasingly popular. In recent years, dense approaches to this problem which uses information from whole input images have achieved compelling results. Because these methods are very different from feature-based approaches, it allows a complementary binding, trying to exploit the advantage of both ways.

Feature-Based Method: The standard approach is to extract a sparse set of salient image features in each image match them in successive frames using invariant feature descriptors. Then robustly recover both camera motion and structure using epipolar geometry. Finally, refine the pose and structure through reprojection error minimization. Lots of algorithm such as [2, 3] follows this procedure. Feature based methods are usually not expensive at computational cost and they can achieve robust real-time camera tracking under high quality corresponding pixels pairs. A sparse point cloud structure usually can be recovered from the tracked feature points.

Direct Method: Alternatively, due to increased computational capabilities, recently monocular direct visual odometry algorithms have been proposed. Direct methods use information from every whole input frame, recovering depth value for each pixel in a selected key-frame. In [1], relatively expensive optimization methods for accurate and fully dense depth maps are realized parallelizedly on GPGPU hardware. Also some researches turned to a semi-dense depth filtering formulation, such as [5, 4], which greatly reduces computational complexity, allowing real-time performance on a CPU. Direct methods are able to reconstruct more surface detail information under statistical framework. Prior knowledge or assumption will be crucial to the result quality.

In this paper, we developed a method of dense 3D reconstruction based on the state of the art algorithm [1] combined with feature-based SLAM. The algorithm of [1] uses passive sensors only so it naturally suffers from ambiguity problem, and sometimes obvious and severe in its reconstruction result for smooth and highly textured surface areas. In [1], the *Hypothesis* that depth discontinuities often coincide with image edges caused those areas appeared to

[*1] Nara Institute of Science and Technology

be potholes-like or stairs-like, obtaining unnecessary and disturbing depth discontinuity. In our research, we carefully analyzed the cause of this phenomenon and related ambiguity problem and come up with a modified regularization process, aiming at reconstructing highly textured surface as a whole smooth block correctly, evaluated by brief experiments in simulated and real environment.

2 Dense Depth Map Construction Using Feature Prior Information

2.1 Cost Volume

The data structure of cost volume is designed as a 3 dimensional matrix to represent the disparity space, which can be simply explained as an array of image size matrices linearly lying on the inverse depth direction, also called disparity. So the cost volume is a $R \times C \times D$ cube matrix, where $R \times C$ is the image size, D is the sampling number along minimum inverse depth ξ_{min} to maximum inverse depth ξ_{max} in the camera view field. A reference frame r consists of a reference image $\boldsymbol{I}_r$ with pose $\boldsymbol{T}_{rw}$ and data cost volume $C_r(\boldsymbol{u}, d)$ collected from the related input frames set $\mathcal{I}$. The cost volume is for storing the intensity difference of each pixel in the reference frame r with any existing pixel lying along its epipolar line in other input frames from set $\mathcal{I}$. For each pixel $\boldsymbol{u}$, a row $C_r(d)$ along the depth direction in the cost volume is computed by projecting a pixel in the reference image where the volume is built, into each of the overlapping images and summing up some similarity measurement, for instance, Huber norm of individual photometric errors $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$. For a whole reference frame, the cost sums up to

$$\boldsymbol{\rho}_r(\boldsymbol{I}_m, \boldsymbol{u}, d) = \boldsymbol{I}_r(\boldsymbol{u}) - \boldsymbol{I}_m(\pi(\boldsymbol{K}T_{mr}\pi^1(\boldsymbol{u}, d))) \quad (1)$$

$$C_r(\boldsymbol{u}, d) = \frac{1}{|\mathcal{I}|}||\boldsymbol{\rho}(\boldsymbol{I}_m, \boldsymbol{u}, d)||, m \in \mathcal{I} \quad (2)$$

The reprojection process is $\pi(\boldsymbol{K}T_{mr}\pi^1(\boldsymbol{u}, d))$, where $\boldsymbol{T}_{mr}$ is the relative transform matrix between camera position of input frame and the reference frame, $\boldsymbol{K}$ is the intrinsic parameter of the camera and π is for dehomogenization. In this paper it is implemented in a 4×4 projection matrix manner where the inverse depth sampling can be integrated in the 3rd row.

2.2 Exploit Feature Points Depth

We first exam the ambiguity from the view of cost volume method to see how the defect brought by *Hypothesis* at section 1 happens. First we suppose a plane separated into two regions with different color. The cost volume method require camera to sweep the scene along a narrow baseline so we imagine the camera is moving from the current position to left or right as figure 1. Figure 2 shows $C_r(d)$ of pixels in different color regions when camera moving along the direction of the arrow in figure 1.

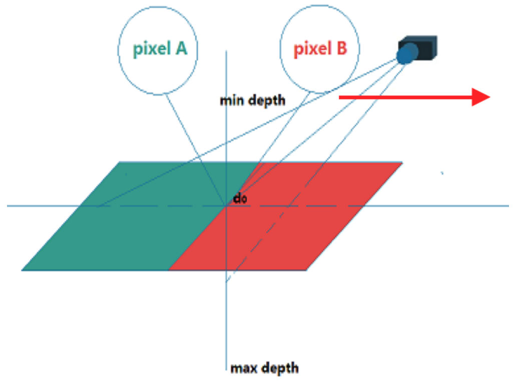

Fig.1 Camera sweeping over a two color plane

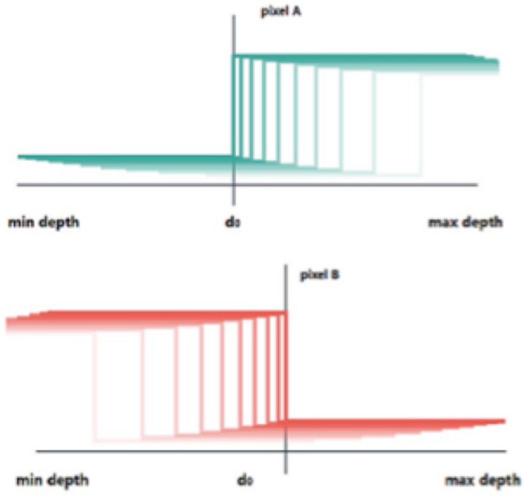

Fig.2 $C_r(d)$ of pixel A and pixel B

2.3 Reconstruction Fixing

This section is about how to fix the accumulated $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$ for feature point, to support the whole optimization process better. The confident depth of feature points is from camera tracking. After performing bundle adjustment integrated with RANSAC, we can track a series of feature points with an relative accurate estimation of their 3D position. We apply the 4×4 projection matrix mentioned in previous section on these feature points, projecting them to the volume. Then compare the minimal in the col-

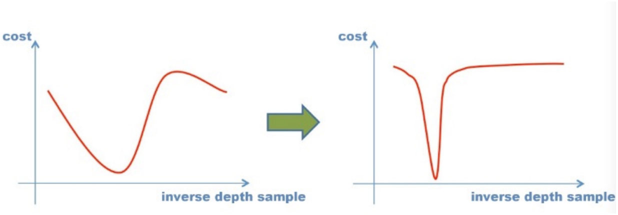

Fig.3 Idea of fixing the confident feature point depth. Left: original $C_r(d)$. Right: $C_r(d)$ with given minimal.

umn of corresponding pixels and the depth interval where this feature point falls. The general spirit of this fixing manipulation is described as figure 3. In detail, for different situations we will fix this column of accumulated $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$ respectively.

1. If the plot of $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$ at targeted position has a minimal which is less than 50% of cost at all the other non-adjacent depth, and this minimal is adjacent to or same with the depth interval where this feature point falls , we do not modify this column.
2. If the plot of $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$ has a minimal which is less than 50% of cost at all the other depth, but it is not adjacent to or same with the depth interval where this feature points falls, we will put the cost at the depth interval where this feature points falls and its adjacent depths to the same value as the minimal.
3. If the plot of $\boldsymbol{\rho}(\boldsymbol{I}_m, d)$ does not have a minimal which is less than 50% of cost at all the other depth, we replace the whole plot into a new curve. The basic spirit of this operation is illustrated as following graph, which the valley part is parabolic.

3 Experiment and Result

3.1 Simulation Analysis

We use Matlab as our simulation experiment environment. The test data set is new Tsukuba stereo data set containing a CG-generated video sequence of 1800 frames with 4 different illumination conditions. The camera trajectory and depth maps for each frames is also provided as ground truth. We applied cost column modification method and anisotropic regularization term based on *Hypothesis* at section 1 simulated environment to confirm the implement and compare the reconstruction results. Figure 4 shows the recovered depth maps for each method. From figure 5 we can see that the optimization processes for two methods start from a same error level about average 16 inverse depth sample intervals and after 380 iterative steps, our method using feature depth as prior information ended up at smaller error level at average 8 inverse depth sample intervals compared with average 9 intervals of anisotropic regularization term method. Figure 6 shows that in the typical area where is smoothed and highly textured, average depth error drops from 25 inverse depth sampling intervals to 9 compared with 14 for anisotropic regularization method after 380 steps iteration .

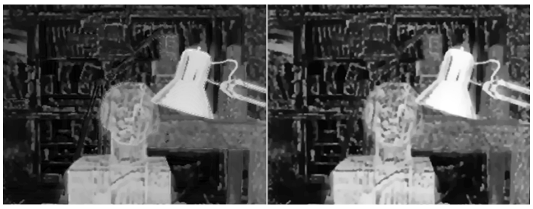

Fig.4 Recovered depth map. Left: cost column fixing; Right: anisotropic regularization

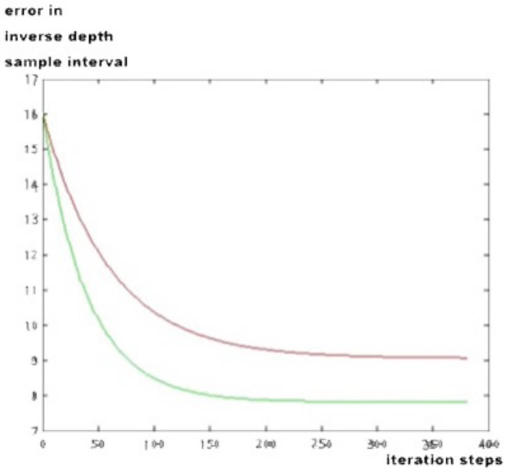

Fig.5 Overall error of depth estimation comparison within 380 iteration, green:cost column fixing, red:anisotropic regularization term

3.2 Real Environment

The test scene is under daylight illumination, consists of a check board, a textured mug, and a blue bendo cloth for background. Material for the mug is very smooth and slightly reflective and specular. Check board of course is the most textured object in this scene, so we can compare how different meth-

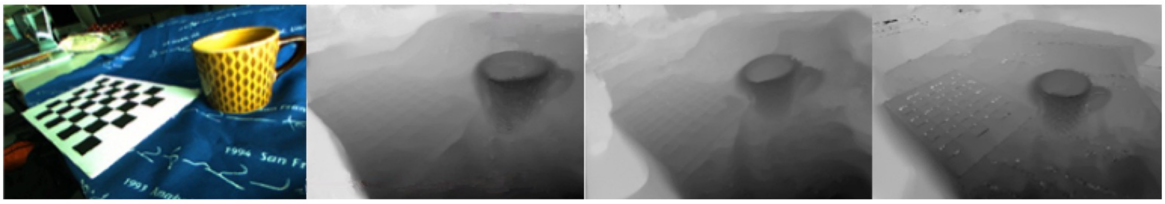

Fig.7 Results in real environment. From left to right:isotropic prior smoothness without fixed feature points depths(most over-fitting);isotropic prior smoothness with fixed feature points depths(proposed method);anisotropic prior smoothness based on *Hypothesis* at section 1

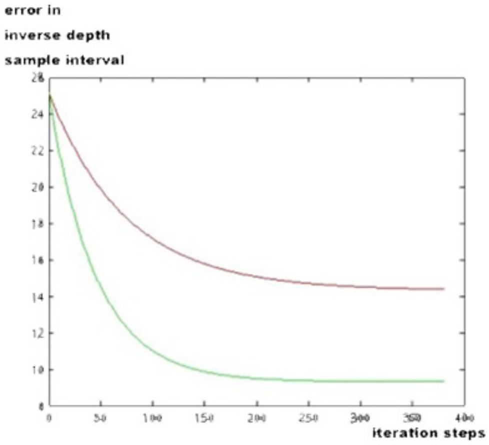

Fig.6 Plot of average error of depth estimation for selected highly textured smooth area within 380 iteration, green:cost column fixing, red:anisotropic regularization term

ods perform on this area mainly focusing on whether they generate stair effects on recovered depth map. The reconstruction of the inner surface of the mug is also where the ambiguity of structure from motion method happens. So we will also compare and analyze the experiment result on these places. From figure 7 we can see the panel-like effect brought by anisotropic regularization term on check board area due to its rich texture and in the result using neither weighted regularization nor depth fixing, the boundary of check board is not clear. We achieved a good compromise between details and over fittings by using feature points depth fixing on isotropic prior smoothness.

4 Conclusion

The proposed method, which uses tracked feature points depth information to modify feature points fixing cost column in the depth map estimation optimization process, was tested in both simulation environment and real environment. In the simulated test we find that our method is slightly lower at the total error with true depth value accumulated over whole frame, and this advantage becomes more clear when compared within highly textured smooth area than the edge based anisotropic regularization method. In the real environment test the proposed method achieves a good compromise between details and over fittings, while avoid the effect brought by original weighted regularization method. For future research direction, finding proper hypothesis for occlusion problem fitting current optimization algorithm is always promising.

References

[1] Richard A. Newcombe and Steven J. Lovegrove and Andrew J. Davison: DTAM: Dense tracking and mapping in real-time; *In Proceedings of International Conference on Computer Vision,* pp. 2320-2327, 2011.

[2] Georg Klein and David Murray: Parallel tracking and mapping for small AR workspaces; *In Proceedings of International symposium of Mixed and Augmented Reality,* pp. 225-234, 2007.

[3] Mingyang Li and Anastasios I. Mourikis: High-precision, consistent EKF-based visual inertial odometry; *International Journal of Robotics Research,* 32, no. 6 pp. 690-711, 2013.

[4] Jakob Engel, Jurgen Sturm, Daniel Cremers: Semi-Dense Visual Odometry for a Monocular Camera; *In Proceedings of International Conference on Computer Vision,* pp. 1449-1456, 2013.

[5] Jakob Engel and Thomas Schöps and Daniel Cremers; LSD-SLAM: Large-Scale Direct Monocular SLAM; *In Proceedings of European Conference on Computer Vision,* pp. 834-849, 2014.