

ROBUST SCENE GEOMETRY RECOVERY FROM EGO-CAMERAS FOR VISUAL NAVIGATION

SUVAM PATRA



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

SEPTEMBER 2019

©Indian Institute of Technology Delhi, New Delhi, 2019

ROBUST SCENE GEOMETRY RECOVERY FROM EGO-CAMERAS FOR VISUAL NAVIGATION

by

SUVAM PATRA

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



Indian Institute of Technology Delhi
September 2019

Certificate

This is to certify that the thesis titled **Robust Scene Geometry Recovery from Ego-cameras for Visual Navigation**, being submitted by **Mr. Suvam Patra** for the award of **Doctor of Philosophy** in Computer Science and Engineering, is a record of bona-fide work carried out by him under our guidance and supervision at the Computer Science and Engineering Department, Indian Institute of Technology Delhi. To the best of our knowledge, the work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Subhashis Banerjee
Professor
Department of Computer Science
and Engineering
Indian Institute of Technology Delhi
New Delhi 110 016

Prem K. Kalra
Professor
Department of Computer Science
and Engineering
Indian Institute of Technology Delhi
New Delhi 110 016

Chetan Arora
Associate Professor
Department of Computer Science
and Engineering
Indian Institute of Technology Delhi
New Delhi 110 016

DEDICATED TO

My Parents and My Teachers.

Acknowledgments

I would like to sincerely thank all my teachers who guided me in my professional and educational growth and motivated me towards research. First and foremost, I would like to express my sincere gratitude towards my supervisors Prof. Subhashis Banerjee, Prof. Prem Kumar Kalra and Prof. Chetan Arora for their continuous guidance. I was fortunate enough to get research directions from them whenever I got stuck. Prof. Banerjee was always enthusiastic and there to help me whenever I needed help. He gave me the freedom to work and explore my problems on my own.

My co-supervisor Prof. Prem Kumar Kalra has always been there to hear me out patiently and help me out with his advice for solving my research problems and also sometimes for my personal development. He also helped me to improve my communication skills and my technical writing.

I would like to sincerely thank my co-supervisor Prof. Chetan Arora who helped me throughout my doctoral journey. His advice and guidance was to the point and precise. It helped me immensely in understanding my research problems and tackle them wisely. His timely responses and suggestions contributed highly to the critical milestones in my research.

I would also like to thank my research committee members Prof. Subodh Kumar, Prof. Naveen Garg and Prof. Sumantra Dutta Roy for their valuable reviews in all my research demonstrations.

I would like to express my gratitude towards Prof. Venu Madhav Govindu of Indian Institute of Science Bangalore for his valuable guidance and expert suggestions throughout my research work.

I would like to thank my dearest friend and my father who always motivated me to pursue higher education and research. I would like to express my gratitude towards my

mother and my brother who were always by my side during my most difficult times.

I would like to specially thank my friend and mentor Brojeshwar Bhowmick for his strong support and guidance which helped me to enhance my knowledge in computer vision. I would like to thank Britty Baby for helping me with my communication and technical writing skills. I would also like to thank my friends and colleagues Syaman-tak Das, Shibashis Guha, Dinesh Khandelwal, Shashank Sharma, Kinshuk Sarabhai, Chinmay Narayan, Gayathri Ananthanarayanan and Ankit Anand for their help and support.

I would also like to thank my friends and lab mates Kartikeya Gupta and Shashank Yadav who helped with my research work; and special thanks to Pranjali Maheshwari for his joint work entitled “A Joint 3D-2D based Method for Free Space Detection on Roads”. I would like to thank the computer science department staff especially Mrs. Rekha Rathi, Mr. Hemant Singh, Mr. Surendra Negi, Mr K. R. Kaushik and Mr. Abhishek Sharma for their assistance.

Last but not the least, I would like to express my immense gratitude towards the almighty for his blessings and grace which helped me throughout my journey.

Suvam Patra

Abstract

In this thesis, we address the challenges in typical motion profiles of ego-cameras and design systems that utilize the camera motion to robustly estimate the location of a wearer or a vehicle and map its environment for navigation. In an ego-centric video, the camera views the same scene multiple times as the wearer’s head sweeps back and forth. We use this specific motion profile to perform short visual loop closures aligned with the wearer’s footsteps and use it to correct the pose inaccuracies arising out of wild camera motions [Patra *et al.*, WACV’17]. Accordingly, we first propose a framework for robust camera pose estimation and use this framework successfully for solving different ego applications such as EgoSampling, Hyperlapse, Gaze Fixation, Temporal Segmentation and Activity Classification where state-of-the-art (SOTA) visual odometry (VO) methods have been reported to fail [1, 2, 3, 4]. This pose estimation framework estimates camera poses relative to their adjacent keyframes and lacks the sense of absoluteness in position and depth necessary for autonomous navigation. We thus improve our pose estimation method to propose a robust structure and pose estimation method for solving both the absolute positions and the environment map of the wearer.

We observe that incremental structure from motion (SfM) algorithms employed in most current VO methods in the presence of unreliable pose and 3D estimates from ego-centric videos often generate drift and eventually lead to failures. We address this issue by stabilizing the camera poses first using 2D techniques such as motion averaging

and then compute 3D structure using bundle adjustment [Patra et al., WACV'19]. Additionally, the use of domain knowledge from camera motion profile (e.g., local loop closures) aids the robustness of the proposed algorithm. We validate the accuracy of the estimated poses on different publicly available VO datasets. Further, visual navigation requires reliable estimates of camera position and structure, and hence we use our algorithm to aid visual navigation using ego-cameras mounted on cars.

The existing free space detection methods for autonomous navigation either involve 2D road segmentation techniques or 3D structure estimation techniques for identifying navigable spaces in front of the vehicle. We initially propose an improved 2D road segmentation method for providing strong road detection priors for unmarked roads, under varying illumination conditions [Yadav and Patra et al., ICIP'17]. The proposed conditional random field (CRF) based method uses SegNet [5] for modelling road texture and color lines model [6] for modeling illumination variations. This method can generalize across datasets and provide better road segmentation when compared to SOTA 2D road segmentation methods. However, 2D road segmentation methods do not provide the depth estimates for detecting navigable free space and fail on cases of uneven textures due to shadows, potholes, road restoration, etc. The use of 3D information becomes necessary to overcome these texture based classification failures. We propose an improved method for free space detection that uses a joint 3D/2D based CRF formulation using the generated higher level 3D road priors from our proposed structure and pose estimation algorithm and 2D priors from SegNet [Patra et al., WACV'18]. Both of these cues complement each other along with illumination invariance from the color lines model to create a comprehensive model for free space detection on roads. Experiments show the superiority of the proposed approach over SOTA. This work contributes to an extensive study on ego-cameras and provides insights for further research on camera pose and structure estimation using challenging motion profiles.

सारांश

इस थीसिस में, हम विशिष्ट गति प्रोफाइल में चुनौतियों का समाधान अहंकार कैमरा के उपयोग से करते हैं और सिस्टम डिजाइन करते हैं कि कैमरे की गति का उपयोग करने के लिए एक पहनने वाले या एक वाहन के स्थान का मजबूती से अनुमान लगाने और इसके नक्शे नेविगेशन के लिए वातावरण प्रस्तुत करते हैं। एक अहंकार केंद्रित वीडियो में, कैमरा विचार एक ही दृश्य कई बार पहनने वाले के सिर वापस आगे और पीछे होने के कारण दिखता है। हम लघु दृश्य पाश प्रदर्शन करने के लिए इस विशिष्ट गति प्रोफाइल का उपयोग करने के लिए लूप क्लोजर के पहनने वाले के नक्शेकदम के साथ गठबंधन किया और इससे जंगली कैमरा गति से उत्पन्न होने वाली अशुद्धियों को हटाया।

तदनुसार, हम पहले मजबूत कैमरे के लिए एक रूपरेखा का प्रस्ताव आकलन मुद्रा और इस ढांचे को हल करने के लिए सफलतापूर्वक उपयोग इस तरह के विभिन्न अहंकार अनुप्रयोगों में करते हैं जहां राज्य के अत्याधुनिक (SOTA) दृश्य ओडोमेट्री (वी.ओ.) तरीकों की सूचना दी गई है विफल होने का। यह मुद्रा आकलन ढांचे का पोसेस के उनके आसन्न कीफरम्स के साथ का रिश्ता के अनुमान करता है। हम इस प्रकार में सुधार एक मजबूत संरचना और मुद्रा का प्रस्ताव करने के लिए हमारी मुद्रा आकलन विधि निरपेक्ष पदों और दोनों को हल करने के लिए आकलन विधि पहनने का पर्यावरण नक्शा करते हैं। हम देखते हैं कि वृद्धिशील संरचना गति से (एसएफएम) एल्गोरिदम में सबसे वर्तमान वी.ओ. तरीकों में कार्यरत अर्बकी वीडियो से अविश्वसनीय मुद्रा और 3 डी अनुमान की उपस्थिति अक्सर बहाव उत्पन्न करते हैं। हम इस पते कैमरा स्थिर करके मुद्रा बन गया है पहले का उपयोग कर 2 डी तकनीक जैसे गति औसत v और फिर बंडल समायोजन का उपयोग कर 3 डी संरचना की गणना करते हैं। साथ ही, से डोमेन ज्ञान का उपयोग कैमरा गति प्रोफाइल की एल्गोरिथ्म की मजबूती के लिए प्रस्तावित करते हैं। हम अनुमानित एल्गोरिथ्म की सटीकता को मान्य विभिन्न सार्वजनिक रूप से उपलब्ध वीओ डेटासेट पर करते हैं। इसके अलावा, दृश्य नेविगेशन कैमरा स्थिति और संरचना के विश्वसनीय अनुमान की आवश्यकता है, और इसलिए हम हमारे एल्गोरिथम का उपयोग गाड़ी के ऊपर स्थित अहंकारी कैमरे के मदद से चलने वाली दिश्य चलन के ऊपर करते हैं।

स्वयक्त नवीगेशन के लिए मौजूदा मुक्त स्तल अन्वेषण तरीके या तो 2 डी विभाजन या फिर 3 डी सहजता अन्वेषण के ऊपर निरफर है । हम शुरू में एक बेहतर 2 डी सड़क का प्रस्ताव के लिए मजबूत सड़क का पता लगाने पूर्व प्रदान करने के लिए विभाजन विधि अचिह्नित सड़कें, अलग-अलग रोशनी की स्थिति में प्रस्तावित सशर्त यादृच्छिक फ्रील्ड (CRF) आधारित विधि के लिए मॉडलिंग सड़क बनावट और रंग लाइनों मॉडल के लिए SegNet का उपयोग करते है । यह विधि भर हर दतासेट के ऊपर सामान्यीकरण कर सकते हैं । और बेहतर सड़क विभाजन तकनीक प्रदान करते हैं । हालाकि 2 डी विभाजन तरीके मुक्त स्तल को ढूंढने के लिए गहराए अनुमान को नहीं प्रदान करता है और असामान बनावटों पे विफल होता है । हम इस के लिए एक बेहतर विधि का प्रस्ताव मुक्त अंतरिक्ष का पता लगाने कि एक संयुक्त 3 डी / 2 डी आधारित सीआरएफ का उपयोग कर तैयार का उपयोग करता है और हमारे प्रस्तावित ढांचे से उत्पन्न उच्च स्तर 3 डी सड़क पूर्व और आकलन एल्गोरिथ्म और सेगनेट पे आधारित तरीके को प्रस्तावित करते हैं । इन संकेतों के दोनों के साथ एक दूसरे के पूरक रंग लाइनों मॉडल से प्रसरण में रोशनी एक बनाने के लिए सड़कों पर मुक्त अंतरिक्ष का पता लगाने के लिए व्यापक मॉडल प्रस्तावित करते हैं । प्रयोग दिखाते हैं की प्रस्तावित दृष्टिकोण की श्रेष्ठता SOTA के ऊपर दिखाते हैं । यह काम अहंकार कैमरा पर एक व्यापक अध्ययन करने के लिए योगदान देता है और कैमरा मुद्रा और संरचना आकलन का उपयोग पर आगे अनुसंधान के लिए चुनौतीपूर्ण गति प्रोफाइल अंतर्दृष्टि प्रदान करता है ।

Contents

Certificate	i
Acknowledgments	iii
Abstract	v
List of Figures	xi
List of Tables	xxi
1 Introduction	1
1.1 3D Geometry and Ego-motion Estimation	3
1.2 Our Contributions	5
1.3 Thesis Outline	7
2 Related Work	9
2.1 Part A: Structure and Camera estimation for ego-cameras	9
2.1.1 Loop Closures in VO	12
2.2 Part B: Free space Detection on Roads	13
2.2.1 2D Techniques	13
2.2.2 3D Techniques	15

3	Ego-camera Motion Estimation for Ego-centric Applications	17
3.1	Background	19
3.1.1	Camera Parameters	19
3.1.2	Pose Estimation	20
3.1.3	Depth Map Estimation	21
3.1.4	Rotation Averaging	22
3.2	Proposed Methodology	23
3.2.1	Short Local Loop Closures Detection	25
3.2.2	Rotation averaging based pose refinement	26
3.2.3	Gauss-Newton Re-initialization	27
3.3	Experiments and Results	28
3.3.1	Ego-centric Applications	31
3.4	Failure Cases	37
4	Robust Structure and Pose Estimation for Ego-cameras	39
4.1	Background	42
4.1.1	Motion Averaging	42
4.1.2	Bundle Adjustment	44
4.2	Proposed Methodology	44
4.2.1	KeyFraming	45
4.2.2	Temporal Window Generation	46
4.2.3	Local Loop Closures	47
4.2.4	Camera Pose Estimation	48
4.2.5	3D Structure Estimation	49
4.2.6	Merging and WBA Refinement with Resectioning	50
4.3	Experiments and Results	50

4.3.1	Ego-centric Videos	52
4.3.2	Vehicle-mounted Cameras	56
4.3.3	Handheld Cameras	59
4.3.4	Relocalization	61
4.3.5	Comparison of our Ego-motion Estimation against our Robust Structure and Camera Pose Estimation Method	63
5	Free Space Modeling: Road Detection for Autonomous Navigation	65
5.1	Background	67
5.1.1	Illumination Invariant Color Modelling	67
5.1.2	Image Segmentation	68
5.2	Proposed Methodology	68
5.2.1	Learning Road Texture	69
5.2.2	Learning Road Color Model	69
5.2.3	CRF Formulation	70
5.3	Experiments and Results	71
6	Free Space Modeling: 3D-2D based Free Space Modeling on roads	75
6.1	Background	76
6.1.1	Structure Estimation	77
6.1.2	Illumination Invariant Color Modeling	78
6.1.3	Image Segmentation	78
6.2	Proposed Methodology	78
6.2.1	Generating Higher Order Cues from the 3D Structure	79
6.2.2	2D-3D Joint Road Detection: Problem Formulation	81
6.2.3	CRF Formulation	81
6.2.4	Free Space in World Coordinates	84

6.3	Experiments and Results	85
6.3.1	Qualitative Results	86
6.3.2	Quantitative Evaluation	89
6.4	Failure Cases	91
7	Conclusion	93
	Bibliography	97
	Publications From the Thesis	111
	Biography	113

List of Figures

1.1	Use of Hololens for an AR application (Source [7]).	2
3.1	Simplified overview of our framework (See Section 3.2 for more details.).	23
3.2	Our framework (KF here denotes keyframes). See Section 3.2 for more details.	24
3.3	Our approach to correction of poses using short loop closures and rotation averaging. (a) shows a typical trajectory where each camera is connected to the next, and due to the large rotation, two cameras (indicated in red) are wrongly estimated. The dotted arrows represent intermediate frames. (b) only keyframes take part in loop closures. (c) shows short loop closure detection using KL divergence and extra matches (dotted edges) are found to close the loops. (d) shows the corrected trajectory (green) after running rotation averaging and one iteration of Gauss-Newton for correcting camera positions. Please zoom in for a better visualization.	25

-
- 3.4 Effect of rotation averaging and Gauss-Newton re-initialization is shown on the Georgia Tech. Social Interaction dataset [8]. The trajectory in blue is before refinement with red ovals depicting breaks in the trajectory due to large rotations (some sample images are also shown), which gets corrected after refinement by our method in pink. The trajectory is depicted for a sequence of 500 frames on which we detected multiple batches of short local loop closures within a time frame of 5-10 seconds. Images are best viewed in color. 27
- 3.5 Comparison of trajectories obtained from ego-centric videos. In each of these images, the trajectories obtained by our method are marked in blue, while those attained by LSD-SLAM are marked in red. Due to the lack of any standardized datasets with ground truth poses for ego-centric videos, the accuracy of the trajectories is demonstrated by overlaying them on GPS data (pink). It should be noted that the LSD-SLAM trajectories shown here are the ones obtained just before the algorithm crashed. The frequent crossed trajectories in LSD-SLAM results is due to noisy 3D estimates maintained by the algorithm which often leads to false matches 30
- 3.6 We show the ego-motion computed using our technique on a video used in [4, 3]. The left and right figures show the computed X and Z unit translation directions respectively. The red curve in the plot shows the frames classified as stationary and the blue curve indicates transit frames. Black curve indicates unused frames in ours as well as the original paper [3]. Please zoom in for a better visualization. 31

-
- 3.7 We use our algorithm to compute ego-motion in a video from [4] to detect ‘static’ and ‘stair-climbing’ activities. Left and right images show X and Y unit translation direction components of the computed ego-motion respectively. The red curve indicates frames classified as stationary; blue indicates stair-climbing and black show frames not used in the original [4] as well as our classification. Please zoom in for a better visualization. 33
- 3.8 Comparing the proposed approach with naive 10× fast forwarding and EgoSampling [1] on a publicly available video [2]. The first row shows output from uniform sampling. The second and third rows show outputs from EgoSampling and proposed approach. Focusing on the location of the tree and the pedestrian reveals that both EgoSampling as well as the proposed approach achieves equivalent stabilization which is much better compared to naive uniform sampling. 34
- 3.9 Wearer’s gaze fixation is easy to detect in an ego-centric video by looking at the constancy of camera look at direction. However, Poleg et al. [3] reported the failure of ego-motion computation and suggested a flow based technique. We use the computed ego-motion from our technique on their videos and successfully detect the gaze fixation instances. The figures show some fixations detected by our method. 36
- 3.10 We tested the proposed method on long shaky sequences from Hyperlapse [2]. The trajectory shown is for the video sequence gl02.mp4 in the Hyperlapse dataset. 37
- 3.11 Our algorithm fails in the above cases. In (a) the wearer is part of multiple environments and (b) suffers from extreme blur. In (c) the scene is formed of non-Lambertian surfaces causing inaccuracies in the pose estimates. (d) has negligible illumination 37

4.1	Incremental nature of SOTA SLAM [9, 10, 11] as well as SfM [12, 13, 14] algorithms are unsuitable for extremely unstable ego-centric video when the pairwise camera pose and 3D estimates are unreliable. We propose a robust structure and pose estimation technique for ego-cameras which solves it as an SfM problem over sliding temporal windows. The SfM problem is solved globally over the window, by first stabilizing poses using rotation and translation averaging, before going for bundle adjustment. The figure shows 3D point clouds and trajectory estimated using the proposed algorithm over 12000 frames from a Hyperlapse [2] <i>bike07</i> sequence, where all the other SLAM and SfM algorithms have been reported to fail. Please zoom in for better visualization.	41
4.2	Overview of our proposed structure and camera pose estimation algorithm (Section 4.2). Please zoom in for better visualization.	45
4.3	Incremental 3D and trajectory estimation is problematic for ego-centric videos due to lack of parallax between successive frames. We propose batch mode processing to stabilize the trajectory estimation first. (a), (b) and (c) show output with a batch size of 1, 30 and 500 respectively. (d) is the reference image. Note that large batch size may also cause problems in motion averaging convergence and breaks in SfM causing trajectory break highlighted in (c), structure error as well as trajectory break, highlighted in (a) but corrected by small batch size in (b). The sequence is taken from Huji dataset [3].	47

4.4	Loop closures are an important step in a SLAM algorithm but may never be applied in an ego-centric video because of natural forward motion of the wearer. Here, we have suggested local loop closures for ego-centric videos. First and second images show structure estimation without and with local loop closures respectively. The third image is the reference view. Note the ‘hanging’ stairs in the first image without loop closure. Please zoom in for better visualization.	49
4.5	Comparison of the estimated structure on a challenging Hyperlapse <i>climbing03</i> sequence [2]. SOTA SLAM algorithms fail here, and authors of hyperlapse have reported using SfM algorithm by manually dividing the sequence into batches of 1400 frames. Our algorithm works without failures on the complete sequence. Left: Dense depth map generated by [2] using CMVS [15]. Middle: Corresponding dense depth map generated by our method. Right: A reference view	52
4.6	Our result on another challenging sequence from the Huji dataset [3]. The wearer is walking in a narrow alley and even makes a sharp 360-degree turn. Left: Estimated trajectory overlaid on Google map. Middle: Dense depth map of a portion obtained using CMVS [15]. Right: Reference view. Please zoom for better visualization.	53
4.7	Ground truth trajectories from GPS corresponding to the three sequences that we captured and used in Table 4.2.	54
4.8	Synthetic experiment setup. The first row shows the synthetic scene from various viewing angles. The second row shows the camera trajectory from the same vantage points corresponding to the scene in the top row. . .	55

-
- 4.9 The three columns show the results obtained from ours, ORB and LSD SLAM respectively. The two rows show 3D point clouds obtained from two different viewing angles. Estimated points are shown in red, whereas blue colored points are corresponding to the ground truth. LSD SLAM is clearly worse. ORB-SLAM point cloud is sparser than us as well and has more RMS error. The histogram of absolute error values (**cm**) in Figure 4.10 indicates that ORB-SLAM error is on an average higher than ours. Please zoom in for a better visualization. 56
- 4.10 The three rows show the histogram of absolute error values (**cm**) corresponding to the three columns in Figure 4.9. Please zoom in for a better visualization. 57
- 4.11 Poor 3D estimation by SOTA is one of the primary reasons for breaks. (a) shows a reference view from a Huji sequence [3], (b) poor structure estimation by ORB-SLAM (road highlighted) just before the break (c) shows the correct structure that we estimated at the same point, which is made dense by CMVS [15] in (d). 58
- 4.12 Breaks and structural inaccuracies by Hierarchical SfM [14] compared with us. (a) shows the perspective view of the 3D reconstruction by Zephyr after PMVS, (b) by us and (c) a reference view. In (d) we show the structural inaccuracy (in red) in the Zephyr reconstruction (top view) where it creates an obstruction in the road in the form of a planar wall and placed all 3d points on it, while our method successfully creates a map of the road with no obstruction (in green). 59

-
- 4.13 We present here the trajectory computed from our method on three sequences from KITTI dataset [16]. Left to right, sequence 03, 04 and 10 from the VO benchmark of the KITTI dataset. Ground truth trajectory is shown in red but is not visible since the computed trajectory is very accurate and almost hides the ground truth trajectory. 61
- 4.14 Left: Dense depth map computed using our method + CMVS [15] on *fr3_str_tex_far* seq. (TUM dataset [17]). Right: Comparison with ground truth trajectory after 7 dof alignment 61
- 4.15 Our pipeline can also use standard feature descriptors for relocalization. The figure shows localized novel cameras on the precomputed trajectory using our method (see text for details). The estimated locations (red dots) near the trajectory indicate successful localization in TUM *fr3_str_tex_far* sequence. 62
- 5.1 We propose to use the color lines model [6], in conjunction with a CNN model in a CRF based framework. The color lines model helps CNN adapt better to varying illumination and road conditions. The proposed model outperforms SOTA on the benchmark as well as the dataset that we captured under the targeted road conditions. 67
- 5.2 Each column depicts results on a different image taken from the KITTI dataset [16]. The first row for each column denotes the output after running Segnet [5] and the second row shows the results by our method. 72
- 5.3 Each column depicts results on a different image that we took on Indian roads. The first row for each column denotes the output after running Segnet [5] and the second row shows the results by our method. 73

5.4	Each column depicts results on a different image taken from the internet. The first row for each column denotes the output after running Segnet [5] and the second row shows the results by our method.	73
6.1	Free road space detection is an important problem for the driving assistance systems. However, (a) 2D image based solutions such as SegNet [5] often fail in the presence of non uniform road texture. (b) On the other hand, methods using 3D point cloud fail to identify fine depth boundaries with the pavement. (c, d) The proposed technique uses both 2D and 3D information to obtain SOTA detection results both in 2D image space (c) and in 3D world coordinates (d).	77
6.2	Overall framework of our method	79
6.3	Road plane detection from the point cloud. Road points are marked in red. The road plane is fitted on the basis of the angle θ between road plane and principal axis and distance d of the principal axis from the ground plane	80
6.4	Our result on a few images from a sequence that we took in the University campus. The first row shows the 2D projection of the free space on the image; the second row shows a depiction of detected free space in 3D on the sparse point cloud obtained using VO.	86
6.5	Our result on some sample images from sequence 02 and 03 of the KITTI VO dataset [16], the free space for each of the images is shown in 2D in the first row(overlaid in pink) and in 3D in the second row as a depiction of detected free space overlapped on the corresponding ground truth depth scan obtained from LIDAR marked in red.	87

6.6	Our result on some sample images from Camvid dataset [18], the free space for each of the images is shown in 2D (overlaid in pink).	87
6.7	Some examples from KITTI VO dataset [16]) where our method can fix errors in free road space detection from only 2D image based priors (first row) or 3D depth based priors (second row). Our results are in the third row (corrected areas marked in green). Please zoom in pdf for a better visualization.	88
6.8	Some examples from Camvid dataset [18] where our method can fix errors in free road space detection from only 2D image based priors (first row) or 3D depth based priors (second row). Our results are in the third row (corrected areas marked in green). Please zoom in pdf for a better visualization.	89
6.9	Failure in free space detection due to inaccuracies in both 2D and 3D inputs. The first image shows segmentation in 2D using SegNet while the second image shows the projection of the inaccurate plane estimated by SLAM which leads to wrong free space estimation (third image). Please zoom in for better visualization.	91

List of Tables

3.1	Performance evaluation metrics	33
4.1	Number of breaks suffered by various methods on 5 videos from the Hyperlapse [2] and the Huji egoseg [3] datasets	53
4.2	Number of breaks suffered by various methods on our three sequences against measured distances	54
4.3	Accuracy analysis of estimated structure and comparison with SOTA using a synthetic scene and different motion profiles.	55
4.4	Our results on videos taken from vehicle-mounted cameras on the KITTI dataset [16]. RMS error of computed trajectories (in meters) with respect to the ground truth trajectory show that we improve upon the SOTA on such videos as well. “X” denotes failure in estimation.	60
4.5	Comparison of RMS error with respect to ground truth trajectory on a few sequences from the TUM dataset [17] of handheld video. Our error is better than LSD SLAM on these sequences and also better than ORB-SLAM and PTAM in most cases. “X” denotes failure in estimation.	62

4.6	Quantitative analysis of relocalization error. We perform relocalization as shown in Figure 4.15 and compute error in camera rotation (degrees) and absolute position (cm) after relocalization for novel frames. Smaller error indicates successful localization.	63
4.7	Comparison of relative rotations and relative translation directions of ego-motion estimation (Chapter 3) against our robust structure and camera pose estimation method (Chapter 4)	64
5.1	Segmentation Results on the KITTI [16] dataset. MaxF: Maximum F1-measure, AP: Average precision as used in PASCAL VOC [19] challenges, PRE: Precision, REC: Recall, FPR: False Positive Rate, FNR: False Negative Rate (the four latter measures are evaluated at the working point MaxF). These acronyms are mentioned in the KITTI Benchmark Suite [16]	72
5.2	Segmentation Results on Camvid [18] Testing dataset	73
6.1	Performance evaluation metrics	89
6.2	Quantitative Results on Camvid [18] dataset	90
6.3	Quantitative Results on KITTI VO [16] dataset	91