Robotic Vision: Understanding Improves the Geometric Accuracy

Javier Civera[1]

***Abstract*— . Paraphrasing Olivier Faugeras in the foreword of [1], making a *robot* see is still an unsolved and challenging task after several decades of research.**

The traditional research has been based on the geometric models of multiple views of a scene, estimating a sparse 3D map of the scene and the camera pose. Recent advances have led to fully dense and real-time 3D reconstructions. Also, there are relevant recent works on the semantic annotation of the 3D maps. This extended abstract summarizes the work of [2], [3], [4], [5], [6] in this direction; in particular using mid and high-level features to improve the accuracy of dense maps.

I. INTRODUCTION

SLAM, standing for Simultaneous Localization and Mapping, aims to estimate from a stream of sensor data a model of the surroundings of the sensor and its egomotion with respect to it.

In the latest decades there has been an intense research on visual SLAM, but its robotic application has been limited by the sparsity of their maps. The traditional –feature-based– techniques rely on the correspondences between image point features; that can only be reliably established for salient image points. [7], [8] are two open-source examples of such feature-based monocular SLAM systems.

Recently, [9], [10], [11], have developed algorithms for real-time, online and dense scene reconstruction from monocular images, opening the doors to a wider applicability of visual SLAM. On the other hand, their maturity is still low. For example, [12] shows that their current accuracy is lower than the one of feature-based techniques.

In our work we improve the accuracy of the standard dense techniques by using mid-level and high-level features. Section II details the dense mapping formulation, sections II-.0.a, II-.0.c and II-.0.b describe the new features, section III shows some experimental results and section IV concludes.

II. DENSE MAPPING

The inverse depth ρ for each pixel u in a reference image is estimated by minimizing the following energy $E(\rho)$

$$E(\rho) = \int \lambda_0 C(u,\rho) + R(u,\rho) + \sum_{\pi=1}^{3} P(u,\rho,\rho_\pi)\partial u \ , \quad (1)$$

$C(u,\rho)$ is the photometric difference of each pixel u backprojected at an inverse depth ρ and projected into several overlapping images. $R(u,\rho)$ is a regularization term –usually the TV-norm. Finally, the three terms in the sum $\sum_{\pi=1}^{3} P(u,\rho,\rho_\pi)$ correspond to the three mid and high-level scene cues. For more details on each term and the optimization of the function the reader is referred to [5].

[1]I3A, Universidad de Zaragoza, Spain {jcivera}@unizar.es

a) SUPERPIXELS (3DS): Superpixels are clusters of pixels that have been segmented based on their color and 2D distance. We will assume that such regions of homogeneous color will be planar. Specifically, we use the superpixel segmentation of [13].

We extract the planes $\Pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_q)$ that fit the superpixels by minimizing a function F of the geometric error ε_k of the reprojected contour of the superpixel k in the r^{th} overlapping frame

$$\hat{\Pi} = \arg\min_{\Pi} \sum_{r=1}^{m} \sum_{k=1}^{q} F(\varepsilon_k) \ . \quad (2)$$

The inverse depth ρ_1 in equation 1 is the intersection of the planes Π with the backprojected ray from the pixel u. For details, see [2], [4].

b) DATA-DRIVEN PRIMITIVES (DDP): A data-driven 3D primitive [14] is a RGB-D pattern learnt from data. The visual part of the primitive should be discriminative enough to be detected on another images, and the depth pattern geometrically consistent.

The depth pattern is modelled by its normals and the RGB pattern by a HOG descriptor and a SVM-based classifier. At detection time, the inverse depth ρ_2 for each pixel is extracted from the primitive normal and the depth from a multiview reconstruction. See [5] for more details.

c) LAYOUT (Lay.): The so-called layout [15] consists on the estimation of the rough geometry of a room and the classification of each pixel u into the classes wall, ceiling, floor and clutter.

We assume that the room is cuboid, so its model is composed of six planes. We estimate their normals using multiview vanishing points and their distances from a geometric reconstruction. From such layout, the inverse depth ρ_3 is computed as the intersection of each pixel with the room boundaries if is is classified as that. If the pixel u is classified as clutter we consider that the depth is not predictable.

III. EXPERIMENTAL RESULTS

Figure 1 shows an illustrative view of our results in some selected sequences from the *NYU* dataset [16]. Notice how close our estimation (6^{th} column) is to the ground truth depth (5^{th} column).

Tables I and II show the median depth error of *DTAM* [11], the sparse feature-based multiview stereo *PMVS* [17] and our algorithm on low-texture and low-parallax sequences respectively –typical failures cases for the geometric estimation. Notice our improvement in every case. Notice also how it comes from different features depending on the sequence,

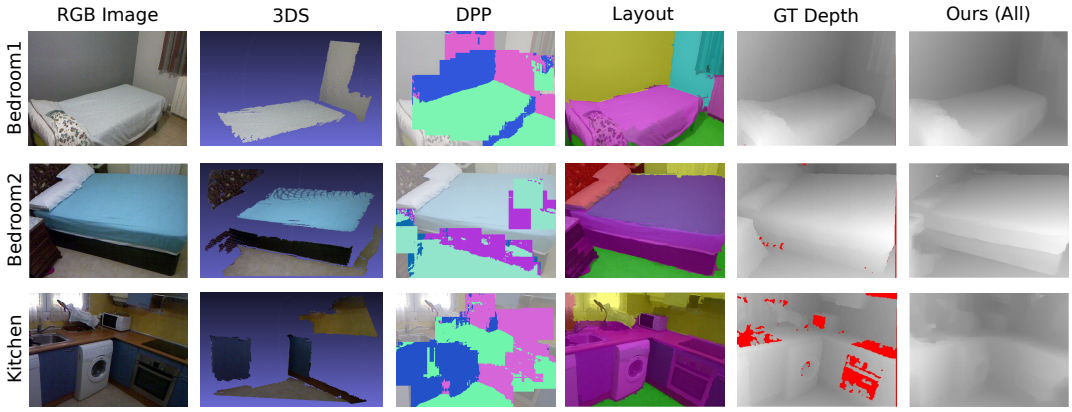

Fig. 1: Estimated depth from 3 sequences –in rows. 1^{st} column is the reference frame. 2^{nd} column are the extracted superpixels, 3^{rd} column the data-driven primitives and 4^{th} column the estimated layout. The 5^{th} column is the ground truth depth from a RGB-D camera and the 6^{th} one our result. Notice the similarity between the latest two.

showing their complementary nature. For more details on these and other experiments see [5].

Sequence	Mean Error[cm]		
	DTAM [11]	*PMVS* [17] (%)	Ours
Bedroom1 (3DS)	15.8	7.0 (18%)	15.0
Bedroom1 (DDS)			**4.2**
Bedroom1 (Lay.)			7.9
Bedroom1 (All)			5.9
Bedroom2 (3DS)	7.1	**5.7** (22%)	**6.7**
Bedroom2 (DDP)			7.6
Bedroom2 (Lay.)			7.7
Bedroom2 (All)			6.8
Kitchen (3DS)	7.2	5.5 (20%)	5.6
Kitchen (DDP)			7.7
Kitchen (Lay.)			5.7
Kitchen (All)			**5.2**

TABLE I: Mean depth error for *DTAM*, *PMVS* and ours in low-texture sequences. (%) is the percentage of pixels reconstructed by *PMVS*.

Sequence	Mean Error[cm]		
	DTAM [11]	*PMVS* [17] (%)	Ours
#1 (Lay.)	9.7	157.5 (3%)	10.4
#1 (DDP)			**7.9**
#1 (All)			9.0
#2 (Lay.)	21.2	43.8 (8%)	8.4
#2 (DDP)			9.2
#2 (All)			**7.6**
#3 (Lay.)	22.2	246.0 (2%)	**12.5**
#3 (DDP)			19.4
#3 (All)			14.5
#4 (Lay.)	42.3	288.4 (9%)	23.8
#4 (DDP)			39.1
#4 (All)			**20.9**

TABLE II: Mean depth error for *DTAM*, *PMVS* and ours in low-parallax sequences. (%) is the percentage of pixels reconstructed by *PMVS*.

IV. CONCLUSIONS

In this abstract –and the associated papers [2], [3], [4], [5], [6]– we have shown how mid and high-level features improve the accuracy of a dense point-based reconstruction from monocular images. The features complement each other nicely, so a fusion of all of them improves the accuracy in a wide array of cases.

ACKNOWLEDGMENTS

This research has been partially funded by projecta DPI2012-32168 and DGA T04-FSE.

REFERENCES

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.
[2] A. Concha and J. Civera, "Using superpixels in monocular SLAM," in *ICRA*, 2014.
[3] A. Concha, W. Hussain, L. Montano, and J. Civera, "Manhattan and piecewise-planar constraints for dense monocular mapping," in *RSS*, 2014.
[4] A. Concha and J. Civera, "DPPTAM: Dense piecewise-planar tracking and mapping from a monocular sequence," in *IROS*, 2015.
[5] A. Concha, W. Hussain, L. Montano, and J. Civera, "Incorporating scene priors to dense monocular mapping," *Autonomous Robots*, vol. 39, no. 3, pp. 279–292, 2015.
[6] M. Salas, W. Hussain, A. Concha, L. Montano, J. Civera, and J. Montiel, "Layout aware visual tracking and mapping," in *IROS*, 2015.
[7] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *ISMAR*, 2007.
[8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós., "ORB-SLAM: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, 2015.
[9] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition*, 2010, pp. 11–20.
[10] G. Graber, T. Pock, and H. Bischof, "Online 3d reconstruction using convex optimization," in *ICCV Workshops*, 2011.
[11] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *ICCV*, 2011.
[12] R. Mur-Artal and J. D. Tardós., "Probabilistic semi-dense mapping from highly accurate feature-based monocular slam," in *RSS*, 2015.
[13] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.
[14] D. F. Fouhey, A. Gupta, and M. Hebert, "Data-driven 3D primitives for single image understanding," in *ICCV*, 2013.
[15] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *ICCV*, 2009.
[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
[17] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.