2013 IEEE International Conference on Robotics and Automation (ICRA)
Karlsruhe, Germany, May 6-10, 2013

Learning depth from appearance for fast one-shot 3-D map initialization in VSLAM systems

Sergio A. Mota-Gutierrez[1], Jean-Bernard Hayet, Salvador Ruiz-Correa,
Rogelio Hasimoto-Beltran and Carlos E. Zubieta-Rico

***Abstract*— The aim of this work is to provide a fast approach for monocular SLAM initialization by constructing an initial 3-D map with interest points that are susceptible to be automatically tracked. Interest points' depth is inferred by means of a linear regression model, which estimates depth on the basis of local image appearance. Our contributions are: (1) a new scheme for learning and predicting associations between depth and local image appearance using RGB-D data; and (2) the use of this scheme for the initialization of state-of-the-art visual SLAM systems from a single image frame. To the best of our knowledge, this is the first attempt to automatically initialize a SLAM system by associating depth to sensor features through machine learning techniques. We performed a series of tests by making use of the celebrated PTAM system and obtained very promising results. We show successful one-shot initialization examples accomplished by applying our proposed approach to unstructured scene environments.**

I. INTRODUCTION

In the last years, monocular simultaneous localization and mapping (SLAM) systems have become very popular in the robotics community as a map building and localization tools. This is mainly because of their low cost, as well as the richness of information that images can convey for solving the data association problem. Since the pioneering work of Davison and collaborators [1], the development of new SLAM algorithms has been a persistent trend in Augmented Reality and Mobile Robotics research. However, all monocular SLAM systems reported in the literature depend critically on specific 3-D map initialization mechanisms due to the projective nature of monocular sensors. The initialization process is required in order to provide an initial distance estimate of particular landmarks in the sensor field of view.

Several methods have been proposed for solving the initialization problem. The leading approach detects known landmarks in the scene, providing information that considerably simplifies the initialization problem. In the absence of both landmarks and known points to be inserted in the visual map, previous solutions typically determine an initial value for the point depth by means of triangulation. This may be implicitly performed by using estimation-oriented schemes such as [1] or explicitly performed in structure from motion-oriented schemes such as [2]. In both cases, several frames may be necessary to accomplish accurate initialization. Moreover, if no prior knowledge is given on the motion of the camera, the reconstruction may only be recovered up to a scale.

[1]All authors are with the Computer Science Department, at the Center for Research in Mathematics (CIMAT), 36240 Guanajuato, GTO., México samota@cimat.mx

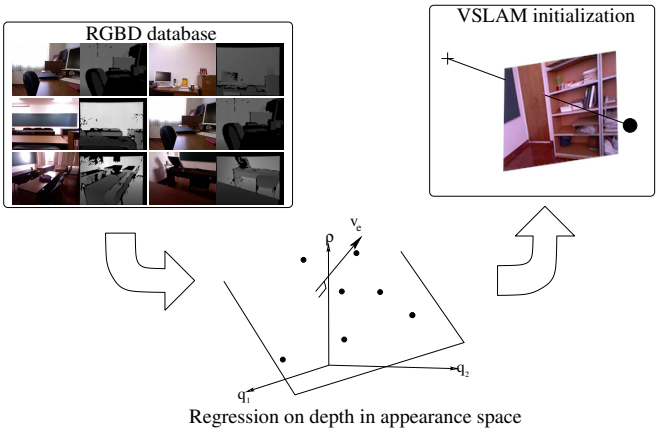


Fig. 1. Overview of our system: We use collected RGB-D data (upper left part) to learn regression parameters in an appearance space (lower part) around corner points to infer depth. The inference system is used to initialize a VSLAM system from a single shot (upper right part).

The paradigm we follow here is based on a key observation. Monocular visual SLAM systems typically operate in environments for which one may have some prior information. Visual reconstruction is often sought in common spaces such as rooms and offices, where visual appearances are not so different from those of places we have already acquired knowledge. Since useful prior information can be readily gathered by means of inexpensive sensors such as RGB-D cameras, designing a prediction scheme that is capable of inferring depth from image appearance is a feasible task.

The aim of this work is to provide a fast approach to initialize a VSLAM system by constructing an initial 3-D map utilizing interest points that are susceptible to be automatically tracked. Interest points' depth is inferred by means of a linear regression model, which estimates depth on the basis of local image appearance (Figure 1).

In summary, our main contributions are: (1) a new scheme for learning and predicting associations between depth and local image appearance using RGB-D data; and (2) the use of this scheme for the initialization of a state-of-the-art visual SLAM system from a single image frame. To the best of our knowledge, this is the first attempt to automatically initialize a SLAM system by associating depth to image features through machine learning techniques.

The reminder of this article is organized as follows. Related work is reported in Section II. The proposed depth-inference strategy that is used for SLAM initialization is described in Section III and the learning of regression

978-1-4673-5643-5/13/$31.00 ©2013 IEEE

parameters in Section IV. We discuss depth inference results in Section V and draw conclusions in Section VI.

II. PREVIOUS WORK

Due to the projective nature of monocular sensors, the initialization problem on visual SLAM is present whenever a new landmark is observed. Early approaches used delayed initialization strategies to solve this problem, where new landmarks are managed by an auxiliary process until they exhibit enough parallax to allow accurate estimations of their degrees of freedom. That is the case of [1], where a uniform distribution over landmarks depth is proposed, and successive observations are used to reduce the depth distribution variance until it can be considered as a low-variance Gaussian.

Delayed initialization strategies do not take advantage of landmark observations until they are fully initialized. This can take a long time when sensor movement is in the direction of landmark location. An approach for undelayed initialization is developed in [3]. This approach maintains multiple depth hypothesis into a Gaussian map for each landmark. Depth hypothesis are spread following a geometric series, and bad hypothesis are pruned by successive observations. In [4], the fact that perspective projection results in a nearly linear image measurement process in inverse depth coordinates has been exploited. This leads to a parametrization that can be used directly into an EKF framework. More of these parametrizations fit to the VSLAM problem, for both points and lines, are presented in [5]. Although these approaches allow for undelayed landmarks initialization, landmark depth is initialized in the form of a widespread distribution.

These approaches have proven useful on robotics and augmented reality applications. However, their applicability is conditioned to the fact of knowing the camera movement between landmark observations. When a scene map is known, camera movement can be estimated from observations. But when there is no map, as when VSLAM applications are started, the sensor motion cannot be estimated. To cope with this problem, the most common strategy has been to assume some a priori knowledge. In [1], a known object is set in the scene and the camera is placed at a specific location relative to this object. In this way, the 3-D locations of the landmarks on the object are known, and constitute the initial map. However, it is not always possible to incorporate known objects to the scene. Moreover, the need for a specific camera pose limits the initialization flexibility. In [2], user intervention is required to tackle this problem. The user is asked to move the camera following a predefined movement between two scene snapshots. Because camera displacement is known, and assuming a planar scene, the camera poses and landmarks locations can be computed. In this approach, the initialization depends on the user ability to perform the predefined movements accurately. Moreover, during the movement, landmarks are tracked by using appearance information only. In order to provide a robust tracking, landmarks are searched over a small image neighborhood. Therefore, smooth movement during initialization is required.

III. INTEREST POINTS DEPTH INFERENCE

Our approach for achieving efficient 3-D initialization of a VSLAM system was motivated by the work of Saxena and colleagues [6]. Their work is aimed at supervised learning for 3-D scene structure recovery in still images of unstructured environments. Broadly, successful learning is achieved by training a Markov random field that encodes spatial relationships between different parts of an image, with local visual features that are highly correlated with depth. Here, we achieve depth learning by means of regularized linear regression over interest points characterized by suitable color and texture features.

Interest points correspond to corner image features that are detected at various scales by means of the FAST algorithm [7]. The i-th interest point of an image is denoted by $\mathbf{p}_i$. Color and texture features associated with $\mathbf{p}_i$ are denoted by the D-dimensional feature vector $\mathbf{q}_i$. These include convolution responses to derivative and Laws' masks operators computed at 3 different scales, leading to feature vectors in $\mathbb{R}^{510}$. Derivative and Laws's operators are designed to characterize texture features from an image [8].

Let ρ_i be the unknown depth of interest point $\mathbf{p}_i$. The goal is to compute an estimate $\hat{\rho}_i$ by modeling image appearance around $\mathbf{p}_i$. To achieve this, we use a linear regression model

$$\hat{\rho}_i = \mathbf{q}_i^T \mathbf{v}_\rho + v_{\rho,0},$$

where $\mathbf{v}_\rho \in \mathbb{R}^D$ and $v_{\rho,0} \in \mathbb{R}$ are the regression parameters to be learned from a training set.

One can expect that this linear model will result in approximate estimations due to large feature vector variability across real data sets. For this reason it is useful to compute an empirical estimate of the error $e_i = \rho_i - \hat{\rho}_i$ from training data. Namely, we use a regression model given by

$$\hat{e}_i = \mathbf{q}_i^T \mathbf{v}_e + v_{e,0},$$

where $\mathbf{v}_e \in \mathbb{R}^D$ and $v_{e,0}$ are the regression parameters. A *confidence measure* ν_i for $\hat{\rho}_i$ can be computed from $\hat{e}_i$ and the *average absolute empirical error* $\bar{\hat{e}}$ as follows

$$\nu_i = 1 - \frac{|\hat{e}_i|}{\bar{\hat{e}} + |\hat{e}_i|}. \tag{1}$$

The key idea of our proposed approach is to initialize a 3-D map using a set of interest points possessing confidence values above a given threshold τ_ν. In practice, a single image may contain a large number of interest points out of which only relatively few possess high confidence values. However, we have observed that as long as the regression model generalizes reasonably well on unseen data, the number of selected points is enough to perform successful initialization. Good generalization performance can be achieved by controlling model complexity and performing training on suitable data sets. The estimation procedure for the $(\mathbf{v}_\rho^T, v_{\rho,0}, \mathbf{v}_e^T, v_{e,0}, \bar{\hat{e}})$ parameters is described in the section that follows. The process for acquiring our training data is described in Section V.

IV. LEARNING REGRESSION PARAMETERS

Training and testing datasets. Our training data set $\mathcal{T}$ consists of a random sample of N $(\mathbf{q}_i, \rho_i)$ pairs. This set is obtained from a database of labeled images by ancestral sampling. Ancestral sampling consists of random sampling of an image from the database, and random sampling of an interest point calculated from that image, from which a feature vector is computed. This process is replicated N times to form $\mathcal{T}$. This set is randomly partitioned into four subsets $\mathcal{T}_i$, $i = 1, \ldots, 4$. The first three are used for training, and $\mathcal{T}_4$ for testing.

Learning parameters. Parameters $\mathbf{v}_\rho$ and $v_{\rho,0}$ are learned by the minimization of the following error function using $\mathcal{T}_1$:

$$\mathbf{v}_\rho^*, v_{\rho,0}^* = \arg\min_{\mathbf{v}_\rho, v_\rho^0} \frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} (\frac{1}{\rho_i}(\mathbf{q}_i^T \mathbf{v}_\rho + v_{\rho,0}) - 1)^2 + \lambda_\rho \|\mathbf{v}_\rho\|_1. \tag{2}$$

Note that $\lambda_\rho \|\mathbf{v}_\rho\|_1$ is a regularization term that controls the complexity of the model and produces sparse regression coefficients. This formulation is the classical LASSO (least absolute shrinkage and selection operator) formulation, which produces a sparse set of regression coefficients, thus enabling automatic feature relevance determination. The interested reader is invited to consult [9] for a detailed description. Implementation details can be found elsewhere [10].

Parameters $\mathbf{v}_e$ and $v_{e,0}$ are learned by the minimization of the following error function using $\mathcal{T}_2$:

$$\mathbf{v}_e^*, v_{e,0}^* = \arg\min_{\mathbf{v}_e, v_{e,0}} \frac{1}{|\mathcal{T}_2|} \sum_{i \in \mathcal{T}_2} (\mathbf{q}_i^T \mathbf{v}_e + v_{e,0} - e_i)^2 + \lambda_e \|\mathbf{v}_e\|_1, \tag{3}$$

where $e_i = \frac{\rho_i - \hat{\rho}_i}{\rho_i} = 1 - \frac{1}{\rho_i}(\mathbf{q}_i^T \mathbf{v}_\rho^* + v_{\rho,0}^*)$ is the fractional error obtained with a depth estimation $\hat{\rho}_i$ vs. training depth ρ_i. Finally, parameter $\bar{\hat{e}}$ is estimated by summing fractional errors over a third subset $\mathcal{T}_3$, as

$$\bar{\hat{e}} = \frac{1}{|\mathcal{T}_3|} \sum_{i \in \mathcal{T}_3} |q_i^T \mathbf{v}_e^* + v_{e,0}^*|. \tag{4}$$

Notice that estimation of the optimal $\mathbf{v}_e^*, v_{e,0}^*$ depends on $\mathbf{v}_\rho^*, v_{\rho,0}^*$, i.e. the errors are inferred under the model built through the first step of the learning phase. Parameter learning in (3) is also done by the method described in [10].

VSLAM initialization. One-shot initialization simply consists of three steps: (1) detecting FAST points in the first image frame provided by the robot/user, and compute their respective descriptors $\mathbf{q}_i$; (2) compute their estimated depths by means of the model $\hat{\rho} = \mathbf{q}_i^T \mathbf{v}_\rho^* + v_{\rho,0}^*$; and (3) compute their estimated errors using $\hat{e} = \mathbf{q}_i^T \mathbf{v}_e^* + v_{e,0}^*$.

Depth and error estimates could be used differently according to the nature of the monocular approach being used. In the case of Kalman-filter-based algorithms, this estimates could allow initial tuning of the Gaussian distributions associated with each map point. In this work, we test our proposed approach in a VSLAM application that uses the well-known PTAM system.

V. RESULTS

This section describes the database used in our experiments, as well as the learning results obtained during the training and testing steps of our proposed approach. It also shows one-shot initialization examples accomplished by applying our approach in unstructured scene environments.

A. Depth/texture database

We acquired RGB images and their corresponding depth maps from indoor environments by means of a Kinect sensor. The sensor is calibrated in order to ensure good correspondences between color and depth maps [11]. Our database consists of 427 RGB-D images, which were captured so that the scene elements are situated within the Kinect sensing range, which is between 800 and 4000 (mm). A few examples of our datasets are shown in Figure 2. Our image database captures a wide range of indoor appearances and depths inside office buildings. We collated a set of 12,810 appearance-depth interest point pairs $(\mathbf{q}_i, \rho_i)$ from our RGB-D images in order to compute a training and testing data sets. Interest points are selected according the procedure described in Section IV. Our training and testing sets consisted of 7686 and 5124 interest point pairs, respectively.

B. Depth estimation experiments

An empirical estimate of the expected value of the actual fractional error in (5), conditioned to the estimated confidence $\hat{\nu}_i$ is greater than a given threshold τ_ν is shown in the plot of Figure 3(a). The shaded area represents the point wise mean plus and minus the estimated 95% confidence interval values. Confidence intervals were computed by means of an approach described in [12]. These calculations were made using data set $\mathcal{T}_4$. This plot shows that high confidence points possess low fractional error values. For instance, points with confidence values higher 80% are expected to have a fractional error e_i around 5%, where e_i is computed as

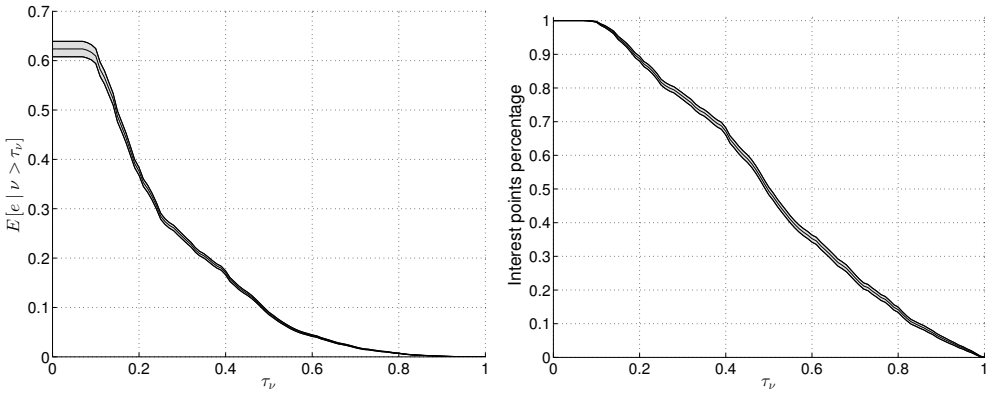
$$e_i = \frac{1}{\rho_i}(\mathbf{q}_i^T \mathbf{v}_\rho^* + v_{\rho,0}^*) - 1. \tag{5}$$

Similarly, in Figure 3(b), we plot the percentage of points for which $\hat{\nu}_i > \tau_\nu$ as a function of τ_ν. The plot shows that about 14.5% of interest points possess a confidence greater than 80%. In general, only relatively few points possess high confidence values, however, we observed in the experiments that a small number of high-confidence points is sufficient to perform fast and successful initialization.

A key observation is that our approach results in high-confidence points across all the images in our database. About 11% percent of the points on each image possesses a confidence greater than 80% (Figure 4). We suggest that our proposed algorithm is generalizing well enough for our application purposes, even-though the distribution of points across images is not completely uniform.

We note that the coefficients $\mathbf{v}_\rho$ and $\mathbf{v}_e$ are relatively sparse, having 73 and 158 non-zero components (out of 510), respectively. The confidence interval for $\bar{\hat{e}}$ is $[.3413, .4164]$ m, and its mean value .3783 m.

Fig. 2. Image examples of our training/testing datasets. The images were collected inside building offices by means of a KINECT sensor.



(a) (b)

Fig. 3. (a) Empirical estimate of the expected value of the fractional error conditioned to the estimated confidence $\hat{\nu}_i$ is greater than a given threshold τ_ν. (b) Percentage of interest points for which $\hat{\nu}_i > \tau_\nu$ as a function of τ_ν.

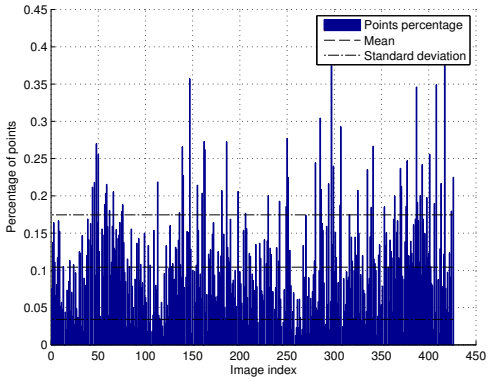


Fig. 4. Percentage of points with in the images of our database that possess confidence values greater than 80%.

C. Initialization of a visual SLAM system

To test the usefulness of our model regarding the initialization of monocular vision systems, we modified the initialization mechanism of the PTAM [2] with our proposed approach. We henceforth refer to the original PTAM initialization as the conventional initialization. In the conventional initialization, the user is asked to perform a camera translation between two snapshots of a planar scene. Given that the camera motion is known and that the scene has planar geometry, both camera poses at the beginning and ending of the motion can be related through an homography. Moreover, 3-D scene point locations can be computed by triangulation.

While this initialization mechanism has been used quite successfully, it is somewhat limited in the sense that it has been designed for use on planar scenes and the scale of the scene can not be recovered accurately (the scale is being given up to the error in the user translation amplitude w.r.t. the suggested motion).

In contrast, our approach not only allows us to perform instantaneous initialization from a single shot, but it can also recover the scene scale, as our method relies on accurate depth learning from a labeled training set. Furthermore, by essence, our approach does not depend on the scene structure, so the planar assumption is unnecessary (Figure 5). We remark that a subset of the interest points used for initialization exhibit persistence along the video sequences as 3-D map elements. This means that their depth estimates were accurate enough to provide an approximately correct initial 3-D map. The reader is referred to the supplementary material for an illustration of the performance of our initialization approach.

We quantitatively compared the conventional and our proposed initialization approach by testing over a set of 20 planar scenes. The Kinect sensor is used to determine a set of ground truth landmarks depths.

The estimated depth for each landmark using the conventional initialization is computed as the magnitude of the vector from the camera center to the estimated landmark position in the scene. Since the estimated depth is known up-

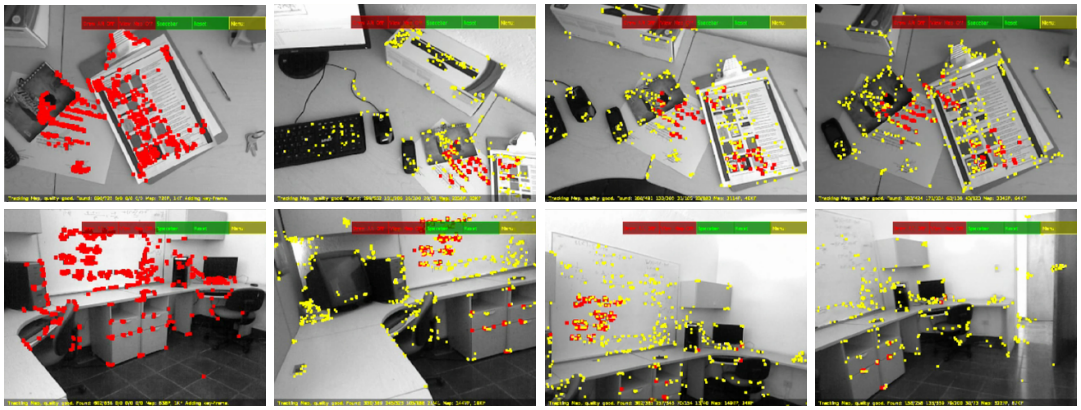


Fig. 5. Our approach allows us to perform instantaneous initialization from a single shot and also recover the scene scale. Our approach does not depend on the scene structure, so the planar assumption is unnecessary. Temporal evolution of the interest points are shown in the sequence of images (from left to right). Red points correspond to interest points obtained with our algorithm. Yellow points correspond to those aggregated by the PTAM engine. Top row: initialization on a planar surface. Bottom row: initialization on a general scene.

to a scale, an additional measurement is performed in order to adjust the estimated depth value of each of the landmarks. The average fractional depth error resulted in a value of 48%.

An empirical estimate of the expected value of the fractional error obtained with our proposed technique, conditioned to the estimated confidence $\hat{\nu}_i$ is greater than a given threshold τ_ν is shown Figure 6(a). The plot indicates that estimates obtained by our approach outperform the conventional initialization regarding landmarks possessing confidence values greater than 65.8%.

Similarly, in Figure 6(b), we plot the percentage of points for which $\hat{\nu}_i > \tau_\nu$ as a function of τ_ν. The plot shows that about 18.2% of interest points possess a confidence greater than 65.8%. This percentage of points was enough to perform successful initialization.

The temporal evolution of the location of the interest points with confidence values above 65% in a typical initialization performed by our approach is shown in Figure 7. We observed that our initial depth estimates are accurate enough to provide the PTAM engine with a suitable 3-D initial map. Illustrations of initial maps are shown in Figure 8. This initial map is modified through the PTAM bundle-adjustment procedure until the original position estimates converge to a stationary value (Figure 7). When this is the case, interest points can be reliably tracked across time and space.

We used video sequences from diverse environments in order to check the generalization ability of our approach as well as the robustness of our depth estimates. We observed that, for indoor scenes our approach allows successful initializations for most of the test cases[1]. Initialization failures occurred mainly on poorly textured images, which are difficult scenarios for approaches based on detection and tracking of interest points. For outdoor scenes the performance of our approach decayed significantly. This is explained by the nature of the images in our database. To overcome this limitation, we re-estimate the model parameters using an outdoor range image dataset [6]. We have obtained promising results under this setting on our preliminary tests.

We found that the learning process was fairly efficient. Convergence was fast and the processing time was adequate. For instance, learning depth from a test data set of 12,000 images takes about 4 min of processing time running Matlab with a 2.5 Gz single Xenon Core. This processing time can be further reduced by using optimized C/C++ and modern computers parallel processing capabilities.

[1]Examples can be found at `http://www.cimat.mx/~samota/projects/VSLAMInitialization.html`

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have described an automatic approach for one-shot initialization of monocular SLAM systems. Our proposed approach is based on depth learning and inference from local image features through standard regularized linear regression. We have shown that our approach results in a set of high-confidence interest points that can be reliably used for construction an initial 3-D visual map. These points tend to be persistent across time and space and therefore can be readily tracked. Several existing issues are guiding our current and future work: a tighter selection of texture descriptors could help us in accelerating the image extraction process; we also aim at better exploiting the nature of interest points, as potential source of depth discontinuities; lastly, we could combine our approach with recursive Bayesian filtering localization and mapping methods, such as the EKF.

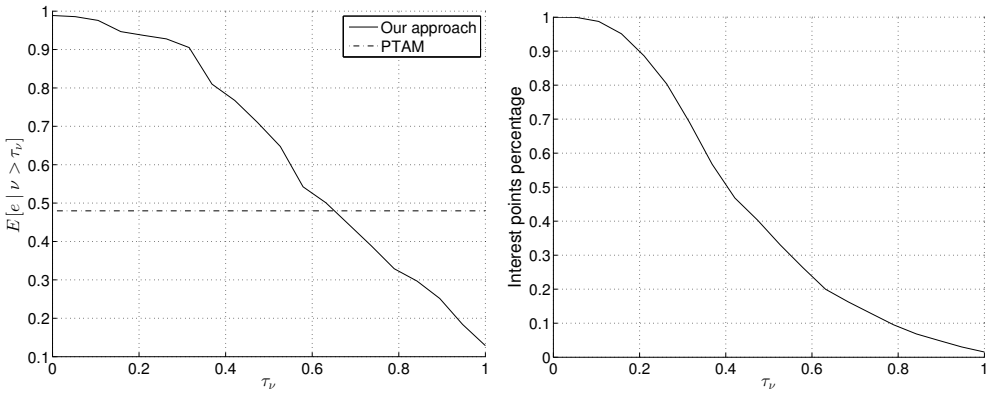
Aknowledgements. We thank G. Klein and his collaborators for providing us with the source code of the PTAM system. This work was partially funded by CONACYT scholarships 253676 and 422831.

REFERENCES

[1] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. of the IEEE Int. Conf. on Computer Vision*, vol. 2, 2003, pp. 1403 – 1410.

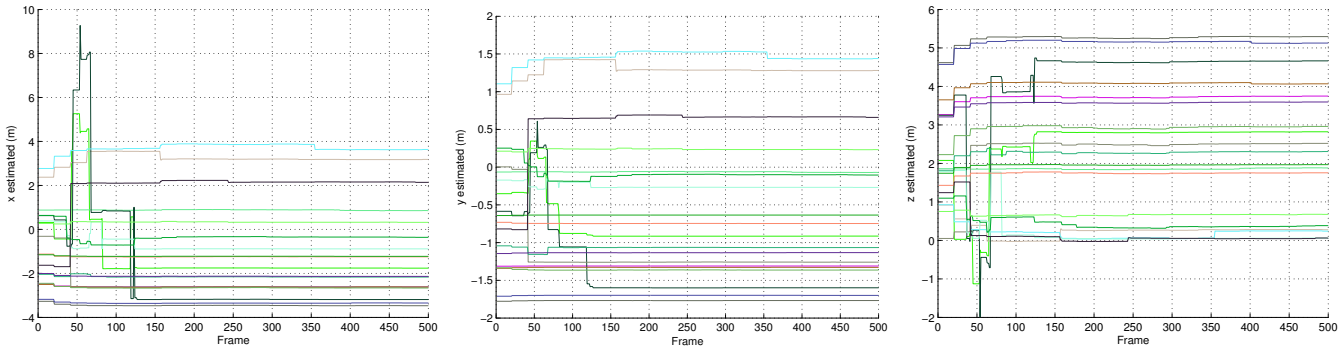
[2] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. of the Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007.

[3] J. Solà, A. Monin, M. Devy, and T. Lemaire, "Undelayed initialization in bearing only slam." in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robot and Systems*, 2005.



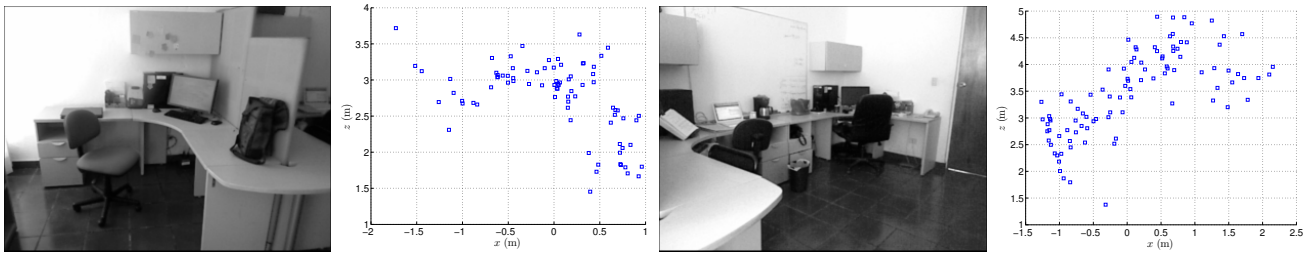
(a) (b)

Fig. 6. Our model as an alternative to PTAM initialization. (a) Empirical estimate of the expected value of the fractional error conditioned to the estimated confidence $\hat{\nu}_i$ is greater than a given threshold τ_ν. (b) Percentage of interest points for which $\hat{\nu}_i > \tau_\nu$ as a function of τ_ν.



(a) (b) (c)

Fig. 7. Spatial-temporal evolution of interest points possessing high confidences. Evolution over time of the 3-D coordinates of the reconstructed landmarks with highest confidences, after a typical initialization. Each curve corresponds to a specific point initially introduced to the map using our method. (a) x-coordinate. (b) y-coordinate. (c) z-coordinate.



(a) (b) (c) (d)

Fig. 8. Initial map, made of high confidence interest points. Note that the scene structure is captured by the point cloud. (a, c) Acquired image. (b, d) Scene map, viewed from a top perspective.

[4] J. Civera, A. J. Davison, and J. M. M. Montiel, "Unified inverse depth parametrization for monocular slam," in *Proc. of Robotics: Science and Systems*, 2006.

[5] J. Solà, T. Vidal-Calleja, J. Civera, and J. M. Montiel, "Impact of landmark parametrization on monocular ekf-slam with points and lines," *Int. J. Comput. Vision*, vol. 97, no. 3, pp. 339–368, May 2012.

[6] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. of Neural Information Processing Systems*, 2005.

[7] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pp. 105–119, 2010.

[8] K. I. Laws, "Textured Image Segmentation," Ph.D. dissertation, University of Southern California, 1980.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[11] C. Herrera, D. Kannala, and J. Juho Heikkila, "Joint depth and color camera calibration with distortion correction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2058–2064, June 2012.

[12] B. Efron, "Bootstrap confidence intervals," *American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987.