Efficiently Increasing Map Density in Visual SLAM Using Planar Features with Adaptive Measurements

José Martínez-Carranza
http://www.cs.bris.ac.uk/~carranza
Andrew Calway
http://www.cs.bris.ac.uk/~andrew

Department of Computer Science
University of Bristol, UK

Abstract

Point based visual SLAM suffers from a trade off between map density and computational efficiency. With too few mapped points, tracking range is restricted and resistance to occlusion is reduced, whilst expanding the map to give dense representation significantly increases computation. We address this by introducing higher order structure into the map using planar features. The parameterisation of structure allows frame by frame adaptation of measurements according to visibility criteria, increasing the map density without increasing computational load. This facilitates robust camera tracking over wide changes in viewpoint at significantly reduced computational cost. Results of real-time experiments with a hand-held camera demonstrate the effectiveness of the approach.

1 Introduction

The visual simultaneous localisation and mapping (SLAM) systems now in widespread use are based on localised point features [3, 4, 8, 9, 10, 11, 14]. Although effective in many respects, the approach has limitations when considering the density and efficiency of map representation. With a dense population of features, camera tracking can be robust, able to withstand significant occlusion and large changes in camera viewpoint. But this comes at a high computational cost, typically increasing quadratically with the number of features.

When building maps over wide areas, the issue can be addressed by using sub-mapping techniques, which are effective mechanisms for containing computational effort as the map grows [1, 7, 12]. However, when building local maps, this is an unsatisfactory and inefficient solution for increasing map density, especially given the likely dependence amongst point features due to physical structure. Sub-mapping in this context merely provides tractability, without addressing the inherent inefficiency of a point based representation.

We propose increasing map density by building in higher-order structure. As model-based tracking systems have demonstrated [15], knowledge of scene structure can yield highly robust and efficient 3-D tracking. One of the main reasons for this is that measurements can be tailored according to camera pose - parameterisation of structure provided by the model allows measurements to be selected at positions and scales according to the predicted view, without any increase in computation. This contrasts with the above SLAM

BMVC 2009 doi:10.5244/C.23.43

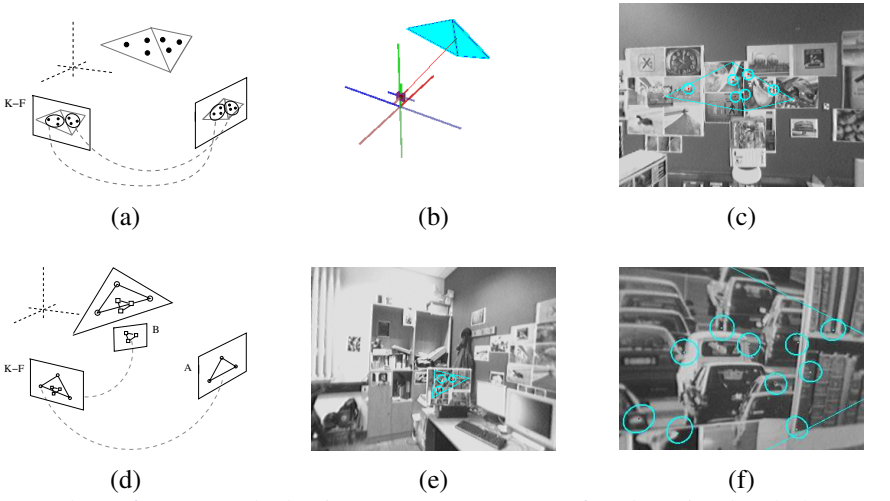


Figure 1: Planar features and adaptive measurement: (a) After detecting local planar structure and inserting planar features into the map, features are updated and used to localise the camera using point based matching with respect to key frames (K-F); (b) A localised camera within a map containing planar features; (c) Point based matches viewed through the camera; (d) Adaptive measurements allow the selection of points for matching according to the camera pose, ensuring that predicted matches lie within the current camera view, whether at a distance (view A) or near to the scene (view B); (e-f) Selected point matches seen through the camera for the far and near cameras.

systems, in which sufficient density of feature points is required to give reliable tracking over a range of views, with the inevitable computational consequences.

Motivated by the above, we augment maps in a Kalman filter based monocular SLAM system with higher-order structure in the form of planar features. These correspond to physical planar patches in the scene and are detected automatically using an appearance method applied to subsets of nearby features in a point based map. Parameterisations are then inserted into the map and maintained with full covariance update. Redundant point features are then removed, giving a hybrid mix of features.

An important and novel aspect of the work is the manner in which the planar features are updated and used to localise the camera. This is based on matching subsets of small region features which lie on the planes and which are adaptively selected according to the current camera pose and to predicted visibility fields for the associated scene patches. These measurement features are defined within key frames linked to the planar features and are projected into the current frame for matching via the parameterised plane. The adaptive selection increases the effective density of the map whilst maintaining a fixed state size in the filter (Figure 1). The result is a computationally efficient system capable of robust localisation and mapping over a wide range of camera views.

1.1 Related Work

We are motivated largely by the work of Gee *et al*. [13]. They demonstrate the incorporation of planar structure into a visual SLAM EKF framework, fitting planes to mapped points and

inserting parameterisations into the map with full covariance update. The work is primarily aimed at detecting large planar structure and the resultant gains from reducing filter state size. However, although it illustrates the utility of using higher-order structure, the method has a number of limitations. Notably, plane detection is based purely on the 3-D estimates of mapped points, allowing the possibility of incorporating non-physical relationships, and the potential for increasing map density is not fully exploited since the measurements are also derived from the previously mapped points. We address both of these issues in this work.

Several other methods have also used planar structure with the aim of improving localisation and mapping in visual SLAM. In point based EKF frameworks, template matching has been improved by utilising local planar patch parameters to enable prior warping, either by on-line estimation of parameters [17] or by assuming fronto-parallel patches at initialisation [11]. Alternatively, planar structure models have allowed the use of image alignment techniques between the current and cached key frames. For example, Silveira *et al*. [19] adopt this approach within an incremental optimisation strategy coupled with a smoothing EKF to give SLAM using only planar measurements. Similarly, Pietzsch [18] uses an iterative EKF framework to estimate planar parameters based on pixel differences between warped templates from key frames. Results reported in all these examples demonstrate the potential gains in terms of robust matching that can be obtained from using planar structure. In this work, we also show this, but also demonstrate how the use of planar models can increase efficiency and map density when working within a point based EKF framework.

2 Overview

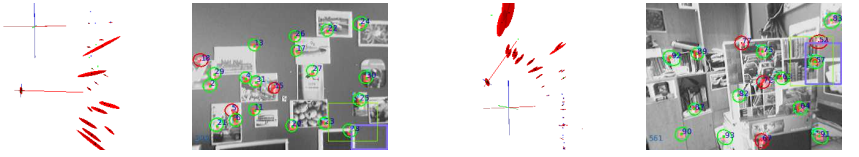
2.1 Point Based Monocular SLAM

We use the extended Kalman filter (EKF) monocular SLAM system developed by Chekhlov *et al*. [4, 10]. This provides real-time estimates of the 3-D pose of a calibrated camera whilst simultaneously mapping the scene in terms of point based features. These are initialised into the map using the inverse depth formulation [6]. For matching we use combined FAST and Shi and Tomasi salient point detection with non-maximal suppression [14] and scale predicted image descriptors [4]. The filter state consists of the camera parameters defined by translation and axis angle rotation vectors plus the respective velocities (12-D), and sets of inverse depth and converged point features (6-D and 3-D, respectively).

During normal tracking, the camera is localised using measurements of mapped features and new features are initialised in regions of the camera view not covered by the map (figure 2). Thus, as the camera explores new areas, the map grows, requiring increasing computational effort in terms of collecting feature measurements and updating the filter state (approximately $O(N^2)$ for N features). When mapping a localised area, if tracking is to be maintained over a wide range of viewpoints, then this map expansion can become computationally inefficient, especially given the likely dependence between close features due to physical scene structure, such as points lying on the same planar surface.

2.2 Incorporating Planar Structure

We address this by introducing higher order structure into the map in the form of planar features. If a portion of the scene is known to be planar and a suitable parameterisation is at hand, then this gives, at least in theory, unlimited map density in that area. In practice,



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

Figure 2: We use an EKF based real-time monocular SLAM system using scale predicted image descriptors for matching. (a) and (c) show the map and localised camera at early and late stages of mapping, with point features and associated uncertainties indicated by ellipses; (b) and (d) show point based matches viewed through the camera with projected uncertainty ellipses (used to constrain matching), with green and red indicating successful and unsuccessful matches, respectively.

it gives greater flexibility in how we choose to derive a measurement for the feature. We can base this on visibility criteria and on the current estimate of the camera pose, aiming to maximise the benefit of the measurements and hence maintain stability of the filter.

For example, in Fig. 1d, a planar feature is updated using a set of point based measurements obtained by matching regions in a key frame (K-F) with those in the current frame (views A or B). These are selected according to the camera pose, ensuring that predicted matches lie within the current view, irrespective of whether the camera is close to the scene (view B, square matches) or at a distance (view A, circular matches). Similarly, estimates of occlusion fields for the feature can also be used to target measurements, helping to minimise the risk of mis-matches (c.f. figure 7). Note also that the knowledge of planar structure helps with the matching itself, enabling the perspective distortion induced by the difference in camera pose to be taken into account.

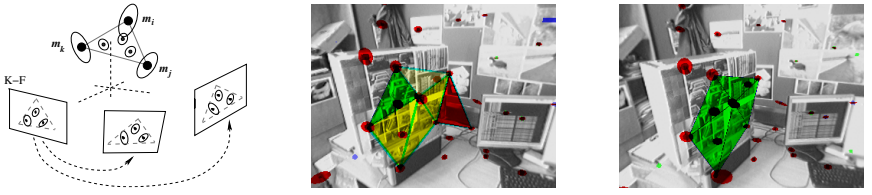
To build this into a real-time monocular SLAM system requires three components: **detection** of planar structure in the scene; **insertion** of planar features into the map; and **adaptive measurement** of the features. Details of these are given in the following sections.

3 Methodology

3.1 Detection of Planar Structure

In order to apply the principle of adaptive measurement it is essential that planar features inserted into the map correspond to actual planar structure in the scene. This contrasts with the approach of Gee *et al*. [13], for example, in which the geometric relationship between mapped points was the main concern so as to reduce state size, rather than the relationship with the scene structure. We require a method for detecting such physical structure in real-time and in tandem with SLAM. For this we employ the method proposed by Martínez-Carranza and Calway [16], which uses an appearance model to detect planes defined by subsets of mapped point features. For completeness we describe the key elements of the method here.

Given a map of point features, we can fit planar structure to subsets of points, either through least squares or triangulation, and then test the hypothesis that a given plane corresponds to physical planar structure in the scene. As illustrated in figure 3, the latter is based on matching salient points lying on the plane across multiple frames using a selected key frame as the reference. In essence, if matching is good over reasonably large changes in



(a) (b) (c)

Figure 3: Appearance based plane detection: (a) The validity of a hypothesised plane derived from the map points $\mathbf{m}_i, \mathbf{m}_j, \mathbf{m}_k$ is tested by matching salient points across multiple frames using the key frame K-F as a reference. This is based on a chi-squared statistics which accounts for the uncertainty in pose and point estimates; (b)-(c) Accepted, possible and rejected planes seen through the camera shown in green, yellow and red, respectively.

pose, then the probability of the hypothesis being correct is high, otherwise the plane can be rejected. An important element for using this approach in SLAM, however, is that the hypothesis test needs to take account of the inherent uncertainty in the estimates of pose and feature positions, as encoded within the covariance of the EKF.

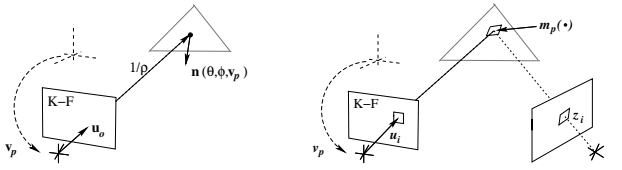
This is achieved in [16] by use of a chi-squared statistic to test whether matches between the sets of salient points over multiple frames are consistent with the estimated uncertainty in the SLAM system. As illustrated in figure 3a, this corresponds to requiring that the matches be within the projected uncertainty bounds derived from the estimated covariance of the mapped points (and hence the plane parameters) and the camera pose. As in [16], we base our plane detection on Delauney triangulation of a point based map, with the caveat that hypothesised planes should have a minimum area. We use small region correlation for matching, with correction for perspective distortion based on the hypothesised plane, about salient points within an appropriate key frame (defined to be the current frame when the plane is first hypothesised).

Examples of plane testing and detection can be seen in figures 3b-c. These show the projected triangulations and accepted, possible and rejected planes in green, yellow and red, respectively. The algorithm has successfully identified the five planes corresponding to actual physical structure. It has also rejected the non-physical plane formed by two points on the foreground box and one from the background wall, as shown in red in figure 3b.

3.2 Insertion of Planar Features

Having detected planar features in the scene these are now inserted into the map using a suitable representation within the filter state. This has two components: a parameterisation of the plane and the current estimate of the camera pose. The latter serves two purposes: it references the plane in the SLAM coordinate system (with the associated uncertainties) and enables subsequent measurement of the planar feature using region based matching with respect to the current frame (key frame). To facilitate the latter the key frame image is also stored in the system. If multiple planar features are initialised within the same frame, then a common key frame and reference pose can be used.

We parameterise in a similar manner to Pietzsch [18] as illustrated in figure 4a. The plane is defined by $\mathbf{y}_p = (\theta, \phi, \rho)$, where (θ, ϕ) defines the unit normal in polar coordinates and ρ is the inverse depth of the plane centre along the ray defined by $\mathbf{u}_o$, with the latter being stored at initialisation of the plane. These are defined within the coordinate system



(a) (b)

Figure 4: Parameterisation and measurement of planar features.

of the current pose $\mathbf{v} = (\mathbf{r}, \omega)$, where $\mathbf{r}$ and the axis angle vector ω define the position and orientation, respectively. Insertion then follows by augmenting the state with a copy of the current pose, i.e. $\mathbf{v}_p = \mathbf{v}$, and with initialised plane parameters $\mathbf{y}_p$ derived from the pose and the subset of mapped points which define the plane, i.e. $\mathbf{m}_i, \mathbf{m}_j, \mathbf{m}_k$ in figure 3.

To allow full covariance update within the filter we also need to adjust the state covariance to take account of the dependency amongst the new and existing parameters. Thus, given an existing state $(\mathbf{v}, \mathbf{m}_1, \ldots, \mathbf{m}_N)$, consisting of camera parameters $\mathbf{v}$ and N point features $\mathbf{m}_i$, we insert the two initialised plane components to give the augmented state $(\mathbf{v}, \mathbf{m}_1, \ldots, \mathbf{m}_N, \mathbf{v}_p, \mathbf{y}_p)$ and then update the state covariance $\mathbf{P}$ according to $\mathbf{P}^{new} = \mathbf{J}\mathbf{P}\mathbf{J}^T$, where the Jacobian $\mathbf{J}$ is given by [1, 13]

$$\mathbf{J} = \left[\begin{array}{ccccccc} \multicolumn{7}{c}{\mathbf{I}_{(M+6)\times(M+6)}} \\ \hline 0 \cdots & \frac{\partial \mathbf{y}_p}{\partial \mathbf{m}_i} & \cdots & \frac{\partial \mathbf{y}_p}{\partial \mathbf{m}_j} & \cdots & \frac{\partial \mathbf{y}_p}{\partial \mathbf{m}_k} & \cdots \quad \frac{\partial \mathbf{y}_p}{\partial \mathbf{v}_p} \end{array}\right] \left[\begin{array}{cccc} \mathbf{I}_{M\times M} & & & \\ \hline \mathbf{I}_{6\times 6} & 0 & \cdots & 0 \end{array}\right] \tag{1}$$

and the partial derivatives are evaluated at the current state estimate. M denotes the original state size and the two matrices defining $\mathbf{J}$ correspond to expansion of the state by each of the two plane components. Multiple and further planar features can be inserted into the state in a similar manner. After the insertion, all the control points are removed from the state as much as from the covariance matrix.

3.3 Adaptive Measurement

Having introduced planar features into the map, they can be updated and used to localise the camera via appropriate measurement. We base this on matching sets of salient points between the key frame and the current frame in a similar manner to that used in the detection process. As illustrated in figure 4b, measurements for a planar feature are therefore assumed to take the following form

$$\mathbf{z}_i = \mathbf{h}(\mathbf{v}, \mathbf{m}_p(\mathbf{v}_p, \mathbf{y}_p, \mathbf{u}_o, \mathbf{u}_i)) + \mathbf{w} \tag{2}$$

where $\mathbf{m}_p(\mathbf{v}_p, \mathbf{y}_p, \mathbf{u}_o, \mathbf{u}_i)$ is the point on the plane which projects to the salient point $\mathbf{u}_i$ in the key frame and $\mathbf{h}$ denotes perspective projection. The noise term $\mathbf{w}$ is assumed to be $N(\mathbf{0}, \mathbf{R})$.

Measurements are obtained by correlating small regions about salient points in the key and current frames, using predicted search regions from the filter measurement covariance and with correction for perspective distortion based on the mean planar structure. These are then used to update the filter state, and hence the plane parameters and camera pose, via the usual EKF equations [2]. Note in particular that given the dependence of the measurement equation in (2) on the key frame pose $\mathbf{v}_p$, this ensures that the latter is also updated via the state covariance within the filter, hence maintaining consistency.

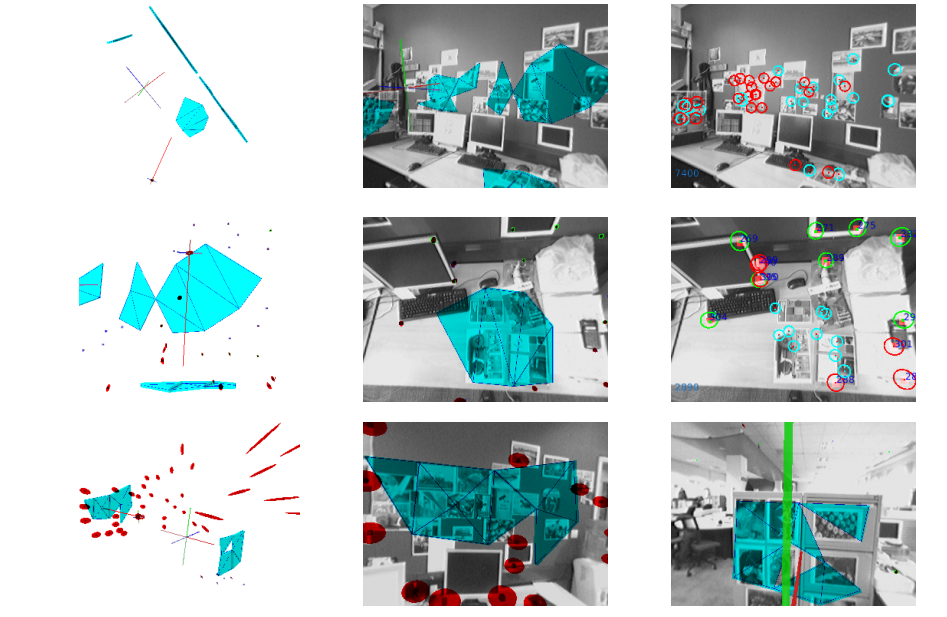
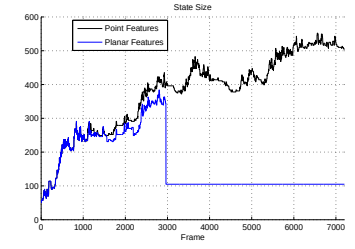


Figure 5: Examples of typical runs of visual SLAM with planar features. Run A - top two rows; run B - bottom row. Images show external views of maps and localised cameras with planar features incorporated and views through the camera of project planar features and selected planar measurements (blue circles).

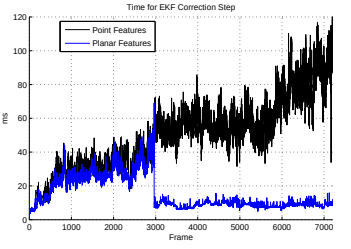
Adaptive selection of the salient points $\mathbf{u}_i$ in the key frame allows the measurements to be tailored according to the current state. We base the selection on two visibility criteria: the points should project into the current camera frame and they should not be occluded. Thus, at initialisation, we identify salient points within the projected area of the feature (the projected Delauney triangle in this work) and at each frame select a subset of points for matching based on these criteria.

Visibility in the current frame follows directly from the mean estimates of the planar structure and two camera poses as illustrated in figure 1d, whilst the probability of occlusion is assessed using an estimated visibility field. The latter is obtained by comparison of pixel values between the projected planar areas in the key and current frames, based on the predicted mean state and normalised w.r.t the respective mean pixel values. High differences in this field indicate potential areas of occlusion and salient points projecting to those areas are then not selected for matching.

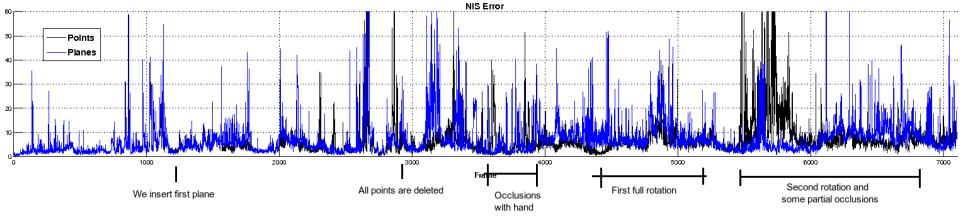
In addition, we can adapt selection according to computational restrictions in terms of limiting the number of measurements that the system can process at any given time. In this case, we select a subset of salient points meeting the above criteria, but with the additional option of replacing points for which the quality of the best match in the current frame is low. The latter provides further refinement of the visibility field and is useful for eliminating bad matches at the occlusion boundaries. As we demonstrate in the following section, this provides an effective mechanism for addressing the trade off between maintaining computational efficiency and stable tracking over a wide range of camera poses.



(a)



(b)



(c)

Figure 6: Comparison of points only and planar feature based visual SLAM in terms of (a) state size, (b) computational cost (EKF update in ms) and (c) normalised innovation squared (NIS) error.

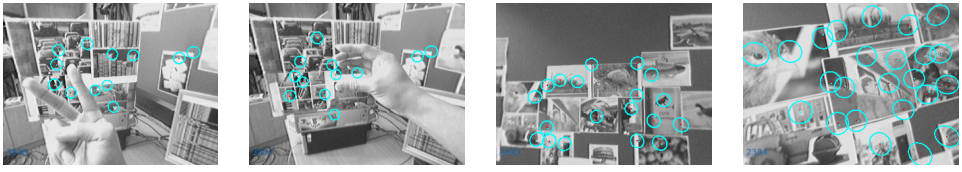
4 Experiments

We tested the performance of visual SLAM with planar features operating in real-time within a laboratory environment using a hand-held web-cam. The camera was calibrated with a resolution of 320x240 pixels and a wide-angled lens with 81^o FOV. All experiments began with building an initial point based map prior to turning on the automatic incorporation of planar features.

Results from two typical runs are shown in figure 5 (runs A and B, top two rows and bottom row, respectively) . This shows external views of the camera and map, with planar features incorporated, and views through the camera showing projected planar triangulations and selected measurement points (blue circles). For run A, a point based map for a relatively small desktop environment was built whilst keeping the camera at approximately the same depth from the scene. Planar feature detection was then turned on whilst the camera explored the map area. Examples of plane detection during this run can be seen in figures 3b-c. Eventually around 20 planar features are detected and successfully inserted into the map. All correspond to physical planes in the scene. Remaining point features were then removed from the map and camera localisation continued using only planar features.

For run B, mapping was carried out over a wider area and planar feature detection was turned on after building a small localised point based map. In this case, planar features are detected and inserted into the map corresponding to two planes in the scene. For the second plane (bottom row, left-hand map in figure 5) this takes place whilst there is high uncertainty about the point based features due to the distance moved by the camera from the start point. Successful incorporation of planar features takes place because these uncertainties are explicitly accounted for in the method, allowing consistent maintenance of the map.

As an illustration of the benefits of using planar features in terms of computational ef-



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

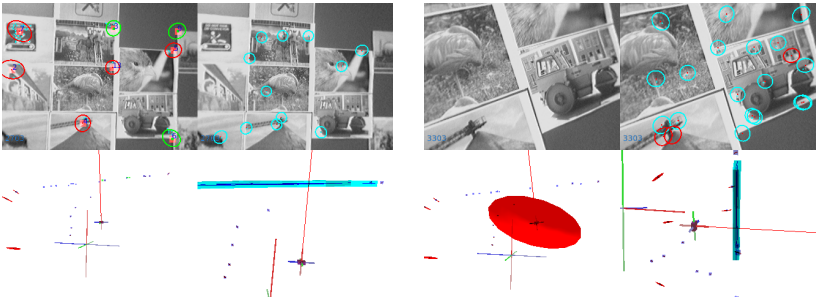
Figure 7: Examples of adaptive measurements: (a)-(b) to avoid occlusion; (c)-(d) to account for changes in camera pose.

ficiency we compared performance in run A with that using only point based features (we ran both methods off-line on the same video). The plots in figure 6 show frame by frame variation of state size, computational time in terms of EKF update and normalised innovation squared (NIS) error [2] for both methods. All point based features were removed from the map at around frame 3000, leaving camera pose estimation to be based only on planar features. Consequently the filter state size is reduced by a factor of around 4 with a corresponding reduction in computational cost. Importantly, when operating with only point based features, the state size continues to grow as the camera explores areas not sufficiently covered by existing point features. This can be contrasted with the use of planar features in which the state remains unchanged and tracking stability is maintained through adaptive measurements, hence illustrating the gain in map density. As indicated in the NIS plots, this advantage is achieved whilst retaining comparable consistency within the filter, except that with planes and adaptive measurement the system is able to provide greater robustness in the face of partial occlusion as can be seen between frames 5500-6000.

The benefits of adaptive measurement are further illustrated in figures 1, 7 and 8. In figures 7a-b, salient point selection ensures that measurements for planar features on the box are positioned to avoid occlusions (in this case between occluding fingers), whilst figures 1e-f and 7c-d illustrate adaptation to take account of camera position, demonstrating the increased map density. The latter is further demonstrated in figure 8. This shows an experiment comparing performance between points only and planar feature maps using a fixed state size, i.e. limiting the map size. In the case of points only mapping, the limited number of features quickly restricts the range for stable camera tracking, resulting in tracking failure as the camera gets too close to the scene. In contrast, with planar features, additional measurements are introduced to maintain tracking without any increase in state size.

5 Conclusions

We have presented a method for using planar features in visual SLAM. The introduction of adaptive measurement is novel and allows exploitation of the potential for increasing map density provided by the incorporation of higher order structure. However, the work presented is at the proof of principle stage. The use of simple Delauney triangulation of the point based map is currently a limitation and we are working towards more sophisticated active search techniques for identifying useful planar structure. Other areas for investigation include the clustering of features that lie on a common planar surface and the use of multiple key frames for a given feature to allow increased tracking range.



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

Figure 8: Comparison of point based and planar based tracking using a fixed state size the state size is fixed. At sufficient distances from the wall, tracking performance is similar (a-b). But as the camera approaches the wall, the density of point features is insufficient to sustain tracking (c), whilst planar parameterisation allows the introduction of new measurements (d), hence maintaining tracking.

Acknowledgements: The authors would like to thank Andrew Gee and Walterio Mayol-Cuevas for valuable discussions and comments. The work was partially supported by CONACYT Mexico under the grant 189903.

References

[1] T. Bailey and H. Durrant-Whyte. Simultaneous localisation and mapping (slam): Part ii - state of the art. *IEEE Robotics and Automation Magazine*, (3), September 2006.

[2] Y. Bar-Shalom, T. Kirubarajan, and X.R Li. *Estimation with Applications to Tracking and Navigation*. 2002.

[3] B.Williams, G.Klein, and I.Reid. Real-time slam relocalisation. In *Proc Int. Conf. Computer Vision*, 2007.

[4] D. Chekhlov, M. Pupilli, W. Mayol-Cuevas, and A. Calway. Real-time and robust monocular slam using predictive multi-resolution descriptors. In *2nd International Symposium on Visual Computing*, November 2006.

[6] J. Civera, A.J. Davison, and J.M.M. Montiel. Inverse depth to depth conversion for monocular slam. In *Proc. Int. Conf. Robotics and Automation*, 2007.

[7] Laura Clemente, Andrew Davison, Ian Reid, José Neira, and Juan Domingo Tardós. Mapping large loops with a single hand-held camera. In *Proc. Robotics: Science and Systems Conference*, June 2007.

[8] A.J Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[9] Andrew J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. Int. Conf. on Computer Vision*, 2003.

[10] D.Chekhlov, W.Mayol-Cuevas, and A.Calway. Appearance based indexing for relocalisation in real-time visual slam. In *Proc. British Machine Vision Conf*, 2008.

[11] E. Eade and T. Drummond. Scalable monocular slam. In *Proc. Int Conf on Computer Vision and Pattern Recognition*, 2006.

[12] Ethan Eade and Tom Drummond. Monocular slam as a graph of coalesced observations. In *Proc. 11th IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.

[13] A.P Gee, D.Chekhlov, A.Calway, and W.Mayol-Cuevas. Discovering higher level structure in visual slam. *IEEE Trans. on Robotics*, 24(5):980–990, 2008.

[14] G.Klein and D.Murray. Parallel tracking and mapping for small ar workspaces. In *Proc. Int. Symp on Mixed and Augmented Reality*, 2007.

[15] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.

[16] J. Martínez-Carranza and A. Calway. Appearance based extraction of planar structure in monocular slam. In *Scandinavian Conference on Image Analysis*, July 2009.

[17] N. Molton, I. Ried, and A.J Davison. Locally planar patch features for real-time structure from motion. In *Proc. British Machine Vision Conf*, 2004.

[18] T. Pietzsch. Planar features for visual slam. In *In Proc. German Conference on Artificial Intelligence (KI 2008)*. Springer, September 2008.

[19] G. F. Silveira, E. Malis, and P. Rives. An efficient direct approach to visual slam. *IEEE Transactions on Robotics*, 24(5):969–979, 2008.