Appearance Based Extraction of Planar Structure in Monocular SLAM

José Martínez-Carranza and Andrew Calway

Department of Computer Science
University of Bristol, UK
{csjmc,csadc}@bristol.ac.uk

Abstract. This paper concerns the building of enhanced scene maps during real-time monocular SLAM. Specifically, we present a novel algorithm for detecting and estimating planar structure in a scene based on both geometric and appearance and information. We adopt a hypothesis testing framework, in which the validity of planar patches within a triangulation of the point based scene map are assessed against an appearance metric. A key contribution is that the metric incorporates the uncertainties available within the SLAM filter through the use of a test statistic assessing error distribution against predicted covariances, hence maintaining a coherent probabilistic formulation. Experimental results indicate that the approach is effective, having good detection and discrimination properties, and leading to convincing planar feature representations[1].

1 Introduction

Several systems now exist which are capable of tracking the 3-D pose of a moving camera in real-time using feature point depth estimation within previously unseen environments. Advances in both structure from motion (SFM) and simultaneous localisation and mapping (SLAM) have enabled both robust and stable tracking over large areas, even with highly agile motion, see e.g. [1,2,3,4,5]. Moreover, effective relocalisation strategies also enable rapid recovery in the event of tracking failure [6,7]. This has opened up the possibility of highly portable and low cost real-time positioning devices for use in a wide range of applications, from robotics to wearable computing and augmented reality.

A key challenge now is to take these systems and extend them to allow real-time extraction of more complex scene structure, beyond the sparse point maps upon which they are currently based. As well as providing enhanced stability and reducing redundancy in representation, deriving richer descriptions of the surrounding environment will significantly expand the potential applications, notably in areas such as augmented reality in which knowledge of scene structure is an important element. However, the computational challenges of inferring both geometric and topological structure in real-time from a single camera are highly

[1] Example videos can be found at http://www.cs.bris.ac.uk/home/carranza/scia09/

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 269–278, 2009.

demanding and will require the development of alternative strategies to those that have formed the basis of current off-line approaches, which in the main are based on optimization over very large numbers of frames.

Most previous work on extending scene descriptions in real-time systems has been done in the context of SLAM. This includes several approaches in which 3-D edge and planar patch features are used for mapping [8,9,10,11]. However, the motivation in these cases was more to do with gaining greater robustness in localisation, rather than extending the utility of the resulting scene maps. More recently, Gee *et al* [12] have demonstrated real-time plane extraction in which planar structure is inferred from the geometry of subsets of mapped point features and then parameterised within the state, allowing simultaneous update alongside existing features. However, the method relies solely on geometric information and thus planes may not correspond to physical scene structure. In [13], Castle *et al* detect the presence of planar objects for which appearance knowledge has been learned *a priori* and then use the known geometric structure to allow insertion of the objects into the map. This gives direct relationship to physical structure but at the expense of prior user interaction.

The work reported in this paper aims to extend these methods. Specifically, we describe a novel approach to detecting and extracting planar structure in previously unseen environments using both geometric and appearance information. The latter provides direct correspondence to physical structure. We adopt a hypothesis testing strategy, in which the validity of planar patch structures derived from triangulation of mapped point features is tested against appearance information within selected frames. Importantly, this is based on a test statistic which compares matching errors against the predicted covariance derived from the SLAM filter, giving a probabilistic formulation which automatically takes account of the inherent uncertainty within the system. Results of experiments indicate that this gives both robust and consistent detection and extraction of planar structure.

2 Monocular SLAM

For completeness we start with an overview of the underlying monocular SLAM system. Such systems are now well documented, see e.g. [14], and thus we present only brief details. They provide estimates of the 3-D pose of a moving camera whilst simultaneously estimating the depth of feature points in the scene. This is based on measurements taken from the video stream captured by the camera and is done in real-time, processing the measurements sequentially as each video frame is captured. Stochastic filtering provides an ideal framework for this and we use the version based on the Kalman filter (KF) [15].

The system state contains the current camera pose $\mathbf{v} = (\mathbf{q}, \mathbf{t})$, defined by position $\mathbf{t}$ and orientation quarternion $\mathbf{q}$, and the positions of M scene points, $\mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M)$. The system is defined by a process and an observation model. The former defines the assumed evolution of the camera pose (we use a constant velocity model), whilst the latter defines the relationship between

the state and the measurements. These are 2-D points $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M)$, assumed to be noisy versions of the projections of a subset of 3-D map points. Both of these models are non-linear and hence the extended KF (EKF) is used to obtain sub-optimal estimates of the state mean and covariance at each time step.

This probabilistic formulation provides a coherent framework for modeling the uncertainties in the system, ensuring the proper maintenance of correlations amongst the estimated parameters. Moreover, the estimated covariances, when projected through the observation model, provide search regions for the locations of the 2-D measurements, aiding the data association task and hence minimising image processing operations. As described below, they also play a key role in the work presented in this paper.

For data association, we use the multi-scale descriptor developed by Chekhlov *et al* [4], combined with a hybrid implementation of FAST and Shi and Tomasi feature detection integrated with non-maximal suppression [5]. The system operates with a calibrated camera and feature points are initialised using the inverse depth formulation [16].

3 Detecting Planar Structure

The central theme of our work is the robust detection and extraction of planar structure in a scene as SLAM progresses. We aim to do so with minimal caching of frames, sequentially processing measurements, and taking into account the uncertainties in the system.

We adopt a hypothesis testing strategy in which we take triplets of mapped points and test the validity of the assertion that the planar patch defined by the points corresponds to a physical plane in the scene. For this we use a metric based on appearance information within the projections of the patches in the camera frames. Note that unlike the problem of detecting planar homographies in uncalibrated images [17], in a SLAM system we have access to estimates of the camera pose and hence can utilise these when testing planar hypotheses.

Consider the case illustrated in Fig. 1, in which the triangular patch defined by the mapped points $\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3\}$ - we refer to these as 'control points' - is projected into two frames. If the patch corresponds to a true plane, then we could test validity simply by comparing pixel values in the two frames after transforming to take account of the relative camera positions and the plane normal. Of course, such an approach is fraught with difficulty: it ignores the uncertainty about our knowledge of the camera motion and the position of the control points, as well as the inherent ambiguity in comparing pixel values caused by lighting effects, lack of texture, etc.

Instead, we base our method on matching salient points within the projected patches and then analysing the deviation of the matches from that predicted by the filter state, taking into account the uncertainty in the estimates. We refer to these as 'test points'. The use of salient points is important since it helps to minimise ambiguity as well as reducing computational load. The algorithm can be summarised as follows:

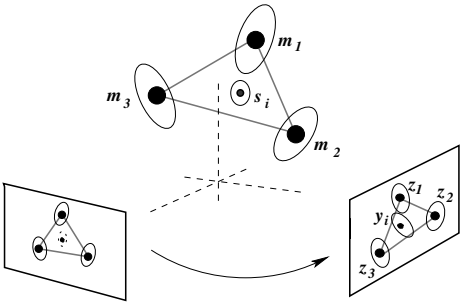

Fig. 1. Detecting planar structure: errors in matching test points y_i are compared with the predicted covariance obtained from those predicted for the control points z_i, hence taking account of estimation uncertainty within the SLAM filter

1. Select a subset of test points within the triangular patch within the reference view;
2. Find matching points within the triangular patches projected into subsequent views;
3. Check that the set of corresponding points are consistent with the planar hypothesis and the estimated uncertainty in camera positions and control points.

For (1), we use the same feature detection as that used for mapping points, whilst for (2) we use warped normalised cross correlation between patches about the test points, where the warp is defined by the mean camera positions and plane orientation. The method for checking correspondence consistency is based on a comparison of matching errors with the predicted covariances using a χ^2 test statistic as described below.

3.1 Consistent Planar Correspondence

Our central idea for detecting planar structure is that if a set of test points do indeed lie on a planar patch in 3-D, then the matching errors we observe in subsequent frames should agree with our uncertainty about the orientation of the patch. We can obtain an approximation for the latter from the uncertainty about the position of the control points derived from covariance estimates within the EKF.

Let $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$ be a set of K test points within the triangular planar patch defined by control points $\mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3)$ (see Fig. 1). From the planarity assumption we have

$$\mathbf{s}_k = \sum_{i=1}^{3} a_{ki} \mathbf{m}_i \tag{1}$$

where the weights a_{ki} define the positions of the points within the patch and $\sum_i a_{ki} = 1$. In the image plane, let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_K)$ denote the perspective projections of the $\mathbf{s}_k$ and then define the following measurement model for the kth test point using linearisation about the mean projection

$$\mathbf{y}_k \approx P(\mathbf{v})\mathbf{s}_k + \mathbf{e}_k \approx \sum_{i=1}^{3} a_{ki}\mathbf{z}_i + \mathbf{e}_k \tag{2}$$

where $P(\mathbf{v})$ is a matrix representing the linearised projection operator defined by the current estimate of the camera pose, $\mathbf{v}$, and $\mathbf{z}_i$ is the projection of the control point $\mathbf{m}_i$. The vectors $\mathbf{e}_k$ represent the expected noise in the matching process and we assume these to be independent with zero mean and covariance R.

Thus we have an expression for the projected test points in terms of the projected control points, and we can obtain a prediction for the covariance of the former in terms of those for the latter, i.e. from (2)

$$C_y = \begin{bmatrix} C_y(1,1) & \cdots & C_y(1,K) \\ \cdots & \cdots & \cdots \\ C_y(K,1) & \cdots & C_y(K,K) \end{bmatrix} \tag{3}$$

in which the block terms $C_y(k,l)$ are 2×2 matrices given by

$$C_y(k,l) = \sum_{i=1}^{3}\sum_{j=1}^{3} a_{ki}a_{lj}C_z(i,j) + \delta_{kl}R \tag{4}$$

where $\delta_{kl} = 1$ for $k = l$ and 0 otherwise, and $C_z(i,j)$ is the 2×2 cross covariance of $\mathbf{z}_i$ and $\mathbf{z}_j$. Note that we can obtain estimates for the latter from the predicted innovation covariance within the EKF [15].

The above covariance indicates how we should expect the matching errors for test points to be distributed under the hypothesis that they lie on the planar patch[2]. We can therefore assess the validity of the hypothesis using the χ^2 test [15]. In a given frame, let $\mathbf{u}$ denote the vector containing the positions of the matches obtained for the set of test points $\mathbf{s}$. Assuming Gaussian statistics, the Mahalanobis distance given by

$$\epsilon = (\mathbf{u} - \mathbf{y})' C_y^{-1} (\mathbf{u} - \mathbf{y}) \tag{5}$$

then has a χ^2 distribution with $2K$ degrees of freedom. Hence, ϵ can be used as a test statistic, and comparing it with an appropriate upper bound allows assessment of the planar hypothesis. In other words, if the distribution of the errors exceeds that of the predicted covariance, then we have grounds based on appearance for concluding that the planar patch does not correspond to a physical plane in the scene. The key contribution here is that the test explicitly and rigorously takes account of the uncertainty within the filter, both in terms of the mapped points and the current estimate of the camera pose. As we show in the experiments, this yields an *adaptive test*, allowing greater variation in matching error of the test points during uncertain operation and tightening up the test when state estimates improve.

[2] Note that by 'matching errors' we refer to the difference in position of the detected matches and those predicted by the hypothesised positions on the planar patch.

We can extend the above to allow assessment of the planar hypothesis over multiple frames by considering the following time-averaged statistic over N frames

$$\bar{\epsilon}_N = \frac{1}{N}\sum_{n=1}^{N} \upsilon(n)' C_y^{-1}(n)\upsilon(n) \tag{6}$$

where $\upsilon(n) = \mathbf{u}(n) - \mathbf{y}(n)$ is the set of matching errors in frame n and $C_y^{-1}(n)$ is the prediction for its covariance derived from the current innovation covariance in the EKF. In this case, the statistic $N\bar{\epsilon}_N$ is χ^2 distributed with $2KN$ degrees of freedom [15]. Note again that this formulation is adaptive, with the predicted covariance, and hence the test statistic, adapting from frame to frame according to the current level of uncertainty. In practice, sufficient parallax between frames is required to gain meaningful measurements, and thus in the experiments we computed the above time averaged statistic at intervals corresponding to approximately 2° degrees of change in camera orientation (the 'parallax interval').

4 Experiments

We evaluated the performance of the method during real-time monocular SLAM in an office environment. A calibrated hand-held web-cam was used with a resolution of 320×240 pixels and a wide-angled lens with 81° FOV. Maps of around 30-40 features were built prior to turning on planar structure detection.

We adopted a simple approach for defining planar patches by computing a Delaunay triangulation [18] over the set of visible mapped features in a given reference frame. The latter was selected by the user at a suitable point. For each patch, we detected salient points within its triangular projection and patches were considered for testing if a sufficient number of points were detected and that they were sufficiently distributed. The back projections of these points onto the 3-D patch were then taken as the test points $\mathbf{s}_k$ and these were then used to compute the weights a_{ki} in (1).

The validity of the planar hypothesis for each patch was then assessed over subsequent frames at parallax intervals using the time averaged test statistic in (6). We set the measurement error covariance R to the same value as that used in the SLAM filter, i.e. isotropic with a variance of 2 pixels. A patch remaining in the 95% upper bound for the test over 15 intervals (corresponding to 30° of parallax) was then accepted as a valid plane, with others being rejected when the statistic exceeded the upper bound. The analysis was then repeated, building up a representation of planar structure in the scene. Note that our emphasis in these experiments was to assess the effectiveness of the planarity test statistic, rather than building complete representations of the scene. Future work will look at more sophisticated ways of both selecting and linking planar patches.

Figure 2 shows examples of detected and rejected patches during a typical run. In this example we used 10 test points for each patch. The first column shows the view through the camera, whilst the other two columns show two different views of the 3-D representation within the system, showing the estimates of camera pose and mapped point features, and the Delaunay triangulations. Covariances

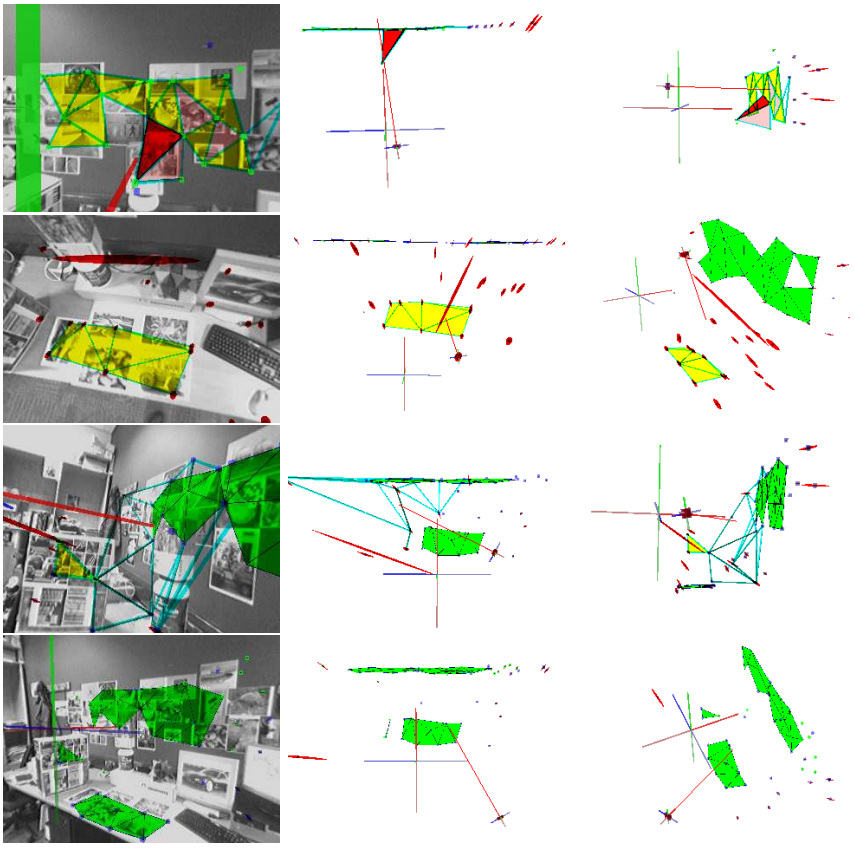

Fig. 2. Examples from a typical run of real time planar structure detection in an office environment: yellow/green patches indicate detected planes; red patches indicate rejected planes; pink patches indicate near rejection. Note that the full video for this example is available via the web link given in the abstract.

for the pose and mapped points are also shown as red ellipsoids. The first row shows the results of testing the statistic after the first parallax interval. Note that only a subset of patches are being tested within the triangulation; those not tested were rejected due to a lack of salient points. The patches in yellow indicate that the test statistic was well below the 95% upper bound, whilst those in red or pink were over or near the upper bound.

As can be seen from the 3-D representations and the image in the second row, the two red patches and the lower pink patch correspond to invalid planes, with vertices on both the background wall and the box on the desk. All three of these are subsequently rejected. The upper pink patch corresponds to a valid plane and this is subsequently accepted. The vast majority of yellow patches correspond to valid planes, the one exception being that below the left-hand red patch, but this is subsequently rejected at later parallax intervals. The other yellow patches are all accepted. Similar comments apply to the remainder of the sequence, with

all the final set of detected patches corresponding to valid physical planes in the scene on the box, desk and wall.

To provide further analysis of the effectiveness of the approach, we considered the test statistics obtained for various scenarios involving both valid and invalid single planar patches during both confident and uncertainty periods of SLAM. We also investigated the significance of using the full covariance formulation in (4 within the test statistic. In particular, we were interested in the role played by the off diagonal block terms, $C_y(k,l)$, $k \neq l$, since their inclusion makes the inversion of C_y computationally more demanding, especially for larger numbers of test points. We therefore compared performance with 3 other formulations for the test covariance: first, keeping only the diagonal block terms; second, setting the latter to the largest covariance of control points, i.e. with the largest determinant; and third, setting it to a constant diagonal matrix with diagonal values of 4. These formulation all assume that the matching errors for the test points will be uncorrelated, with the last version also making the further simplification that they will be isotropically bounded with a (arbitrarily fixed) variance of 4 pixels. We refer to these formulations as *block diagonal 1*, *block diagonal 2* and *block diagonal fixed*, respectively.

The first and second columns of Fig. 3 show the 3-D representation and view through the camera for both high certainty (top two rows) and low certainty (bottom two rows) estimation of camera motion. The top two cases show both a valid and invalid plane, whilst the bottom two cases show a single valid and invalid plane, respectively. The third column shows the variation of the time averaged test statistic over frames for each of the four formulations of the test point covariance and for both the valid and invalid patches. The forth column shows the variation using the full covariance with 5, 10 and 20 test points. The 95% upper bound on the test statistic is also shown on each graph (note that this varies with frame as we are using the time averaged statistic).

The key point to note from these results is that the full covariance method performs as expected for all cases. It remains approximately constant and well below the upper bound for valid planes and rises quickly above the bound for invalid planes. Note in particular that its performance is not adversely affected by uncertainty in the filter estimates. This is in contrast to the other formulations, which, for example, rise quickly with increasing parallax in the case of the valid plane being viewed with low certainty (3rd row). Thus, with these formulations, the valid plane would eventually be rejected. Note also that the full covariance method has higher sensitivity to invalid planes, correctly rejecting them at lower parallax than all the other formulations. This confirms the important role played by the cross terms, which encode the correlations amongst the test points. Note also that the full covariance method performs well even for smaller numbers of test points. The notable difference is a slight reduction in sensitivity to invalid planes when using fewer points (3rd row, right). This indicates a trade off between sensitivity and the computational cost involved in computing the inverse covariance. In practice, we found that the use of 10 points was a good compromise.

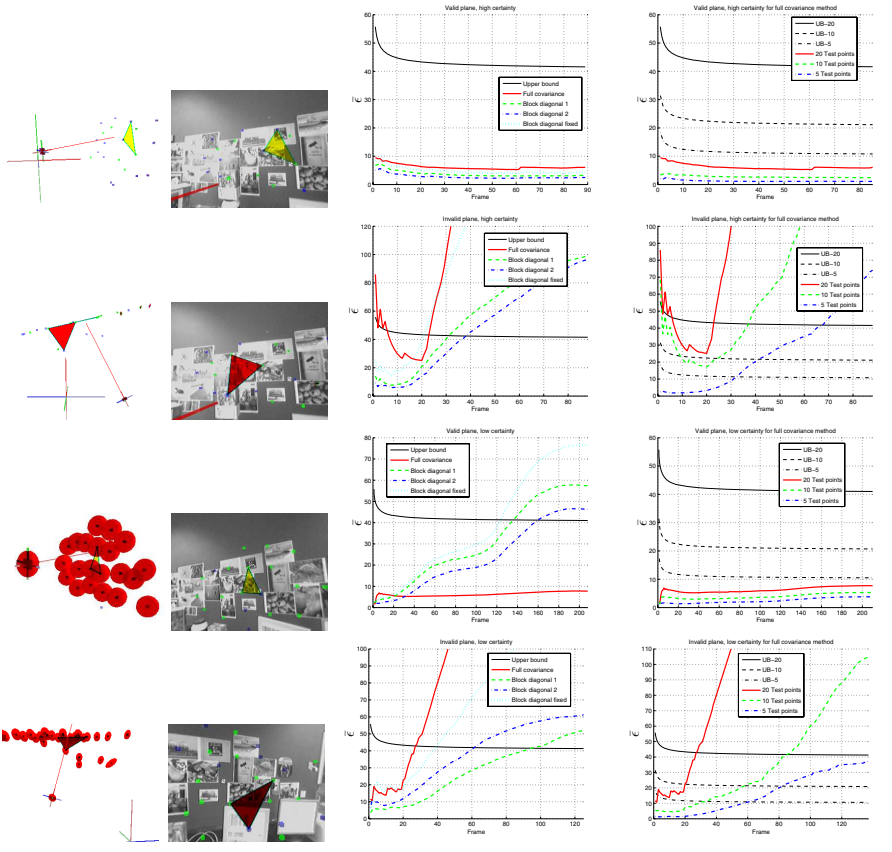

Fig. 3. Variation of the time averaged test statistic over frames for cases of valid and invalid planes during high and low certainty operation of the SLAM filter

5 Conclusions

We have presented a novel method that uses appearance information to validate planar structure hypotheses in a monocular SLAM system using a full probabilistic approach. The key contribution is that the statistic underlying the hypothesis test adapts to the uncertainty in camera pose and depth estimation within the system, giving reliable assessment of valid and invalid planar structure even in conditions of high uncertainty. Our future work will look at more sophisticated methods of selecting and combining planar patches, with a view to building more complete scene representations. We also intend to investigate the use of the resulting planar patches to gain greater stability in SLAM, as advocated in [12] and [19].

Acknowledgements. This work was funded by CONACYT Mexico under the grant 189903.

References

1. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: Proc. Int. Conf. on Computer Vision (2003)
2. Nister, D.: Preemptive ransac for live structure and motion estimation. Machine Vision and Applications 16(5), 321–329 (2005)
3. Eade, E., Drummond, T.: Scalable monocular slam. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (2006)
4. Chekhlov, D., Pupilli, M., Mayol-Cuevas, W., Calway, A.: Real-time and robust monocular SLAM using predictive multi-resolution descriptors. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Nefian, A., Meenakshisundaram, G., Pascucci, V., Zara, J., Molineros, J., Theisel, H., Malzbender, T. (eds.) ISVC 2006. LNCS, vol. 4292, pp. 276–285. Springer, Heidelberg (2006)
5. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Proc. Int. Symp. on Mixed and Augmented Reality (2007)
6. Williams, B., Smith, P., Reid, I.: Automatic relocalisation for a single-camera simultaneous localisation and mapping system. In: Proc. IEEE Int. Conf. Robotics and Automation (2007)
7. Chekhlov, D., Mayol-Cuevas, W., Calway, A.: Appearance based indexing for relocalisation in real-time visual slam. In: Proc. British Machine Vision Conf. (2008)
8. Molton, N., Ried, I., Davison, A.: Locally planar patch features for real-time structure from motion. In: Proc. British Machine Vision Conf. (2004)
9. Gee, A., Mayol-Cuevas, W.: Real-time model-based slam using line segments. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Nefian, A., Meenakshisundaram, G., Pascucci, V., Zara, J., Molineros, J., Theisel, H., Malzbender, T. (eds.) ISVC 2006. LNCS, vol. 4292, pp. 354–363. Springer, Heidelberg (2006)
10. Smith, P., Reid, I., Davison, A.: Real-time monocular slam with straight lines. In: Proc. British Machine Vision Conf. (2006)
11. Eade, E., Drummond, T.: Edge landmarks in monocular slam. In: Proc. British Machine Vision Conf. (2006)
12. Gee, A., Chekhlov, D., Calway, A., Mayol-Cuevas, W.: Discovering higher level structure in visual slam. IEEE Trans. on Robotics 24(5), 980–990 (2008)
13. Castle, R.O., Gawley, D.J., Klein, G., Murray, D.W.: Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In: Proc. Int. Conf. Robotics and Automation (2007)
14. Davison, A., Reid, I., Molton, N., Stasse, O.: Monoslam: Real-time single camera slam. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(6), 1052–1067 (2007)
15. Bar-Shalom, Y., Kirubarajan, T., Li, X.: Estimation with Applications to Tracking and Navigation (2002)
16. Civera, J., Davison, A., Montiel, J.: Inverse depth to depth conversion for monocular slam. In: Proc. Int. Conf. Robotics and Automation (2007)
17. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000)
18. Renka, R.J.: Algorithm 772: Stripack: Delaunay triangulation and voronoi diagram on the surface of a sphere. In: ACM Trans. Math. Softw., vol. 23, pp. 416–434. ACM, New York (1997)
19. Pietzsch, T.: Planar features for visual slam. In: Dengel, A.R., Berns, K., Breuel, T.M., Bomarius, F., Roth-Berghofer, T.R. (eds.) KI 2008. LNCS, vol. 5243. Springer, Heidelberg (2008)