High-level scene structure using visibility and occlusion

Paul McIlroy
http://mi.eng.cam.ac.uk/~pmm33
Roberto Cipolla
http://mi.eng.cam.ac.uk/~cipolla
Ed Rosten
http://mi.eng.cam.ac.uk/~er258

Machine Intelligence Laboratory
Department of Engineering
University of Cambridge
Cambridge, UK

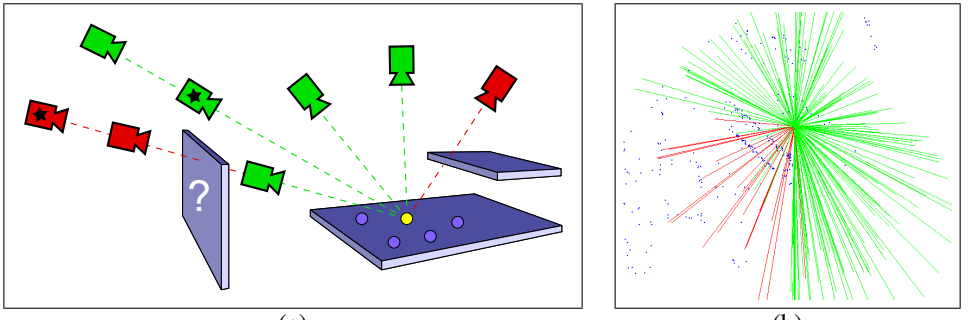
Abstract

We demonstrate a new method for extracting high-level scene information from the type of data available from simultaneous localisation and mapping systems. We model the scene with a collection of primitives (such as bounded planes), and make explicit use of both visible and occluded points in order to refine the model. Since our formulation allows for different kinds of primitives and an arbitrary number of each, we use Bayesian model evidence to compare very different models on an even footing. Additionally, by making use of Bayesian techniques we can also avoid explicitly finding the optimal assignment of map landmarks to primitives. The results show that explicit reasoning about occlusion improves model accuracy and yields models which are suitable for aiding data association.

1 Introduction

The state of the art in structure from motion is advanced, producing accurate point clouds that rival those generated by laser range scans. Meshes generated from point clouds may be refined by dense stereo methods to produce detailed surface meshes. Recently, this pipeline has been demonstrated operating in real-time to produce a live dense reconstruction of a cluttered desktop environment [12]. The live reconstruction pipeline begins with an online structure from motion or visual SLAM system to generate a sparse SFM point cloud, in this case Klein and Murray's PTAM [9] for small AR workspaces is used. Scenes with significant occlusion challenge the state of the art in online SFM and visual SLAM. Wangsiripitak and Murray [16] investigate the difficultly posed by occlusion and propose techniques for dealing with this problem including using higher order scene structure to anticipate occlusion. In this work we propose a model-based approach that exploits both visibility and occlusion information to recover high order scene structure. The use of visibility information to remove erroneous structure by space carving is well-established, Pan *et al.* [13] demonstrate a system using this technique for online reconstruction. In this work we exploit occlusion information, in addition to visibility data, to provide evidence for missing structure.

We consider the visibility and occlusion of map landmarks with respect to camera poses from a landmark centric perspective, with rays emitting radially from landmarks to camera



(a) (b)

Figure 1: Visibility and occlusion: (a) Successful (green) and unsuccessful (red) observation of a landmark (yellow) provides evidence for unmodeled structure causing occlusion (plane marked with '?'). Cameras marked with ⋆ provide redundant information about occlusion.; (b) the view sphere for a single landmark in a desk scene with occlusion.

centres. The dual structure of the projective reconstruction problem with respect to camera centres and 3D point positions has been known since the work of Carlsson [1] and Hartley [7]. More recently Pirker *et al.* [14] exploited camera orientations viewed relative to each individual map landmark to aid data association and map maintenance. In this paper we propose a method for efficiently storing visibility and occlusion information from cameras with respect to each landmark that reduces redundancy in the data whilst maintaining salient information.

Attempts to exploit higher order structure in online SFM and visual SLAM systems have been the focus of recent research. Gee *et al.* [6] proposed a method for folding lines and planes into the state space of an EKF-SLAM system. Carranza and Calway [2] demonstrated that including planar features increases the map density for little additional computational cost and proposed a parameterisation [3] that unifies the treatment of points and planes at initialisation.

Prankl *et al.* [15] propose a method, based on the Minimum Description Length criterion, to control model complexity in terms of the number of planes used to describe the model and prevent over-fitting. They assign a fixed cost to each plane model added. A plane proposal is only accepted if the saving in terms of a cost function on map point distance perpendicular to the plane exceeds the fixed cost per plane. In our work we extend the incremental model selection approach to include visibility and occlusion information. We cast the problem in the Bayesian model evidence framework. With this approach it is possible to compare vastly different models with different modality and dimensionality on an equal footing without assigning fixed costs to the competing models. The merit of this approach has been demonstrated in an architectural setting for the semantic reconstruction of building facades [5]. In this work we describe scene structure using generic model components that describe both planar surfaces and 3D structures. The result is a scene structure hypothesis that describes both the 3D geometry and the visibility/occlusion information from each camera.

The contributions of this paper are two-fold. In Section 2 we propose a method that allows the effective extraction of visibility and occlusion information with low redundancy. In Section 3 we describe a Bayesian model evidence based approach to recover the scene structure compatible with the salient visibility and occlusion data from Section 2. Section 4 provides implementation details describing view sphere generation, model parameterisation and the optimization of the scene structure hypothesis.

2 Visibility and occlusion

A visual SLAM system operates by matching landmarks in a map into images from various viewpoints. The success or failure of the matching contains some information about the visibility or occlusion of that landmark from those views. This is illustrated in Figure 1. Each landmark is projected into every frame in which it is visible, and a estimate of the probability of occlusion is generated by matching the landmark's descriptor to the image.

For a visual SLAM system operating at frame-rate the occlusion information provided by each camera pose is highly correlated with the previous frame and therefore much of the occlusion information is redundant. Not only does the quantity of data grow with $O(KN)$ for K cameras and N map points, but the redundancy itself presents a challenge for methods that aim to exploit the occlusion information. The data is far from i.i.d. and it would therefore be necessary to consider the full joint distribution describing the relationship between the data points to prevent bias towards information from frames repeatedly viewed by a stationary camera.

The redundancy in visibility and occlusion is illustrated in Figure 1a. The red camera marked with $\star$ indicates that the landmark is occluded by unmodeled structure. However, this information is redundant given the presence of the closer camera. The same principle applies, but in reverse for cameras which successfully observe points.

For view rays emitted radially from a given landmark, the salient points along each view ray are the furthest camera centre in which the point is visible and the nearest camera that is occluded. These two points provide evidence for occluding scene structure lying along the ray between the last visible and first occluded camera centres.

2.1 The view sphere

We propose a method that tackles redundancy in visibility and occlusion information by considering a viewing sphere around each landmark. All rays from cameras in which the landmark could be visible to the camera are considered. The rays are grouped into two dimensional bins on the surface of the sphere. Within each bin, the furthest camera in which the point is visible and the closest camera in which the point is occluded are retained. The view sphere provides an upper bound on the number of camera views stored for each map point, at most two per bin. The complexity of the visibility and occlusion is therefore reduced to linear in the map size.

Binning is necessary as no two camera centre to map point pairs are exactly collinear. Binning also reduces the coupling of observations due to consecutive camera frames by a sparse sampling of the view sphere to the extent that occlusion/visibility information can be considered i.i.d. The optimum bin size is a trade-off between between redundancy in the occlusion data and the loss of useful information corresponding to fine structure. A finely quantized view sphere is depicted in Figure 1b for a single landmark in a desk scene with occlusion.

By comparison, Pirker *et al.* [14] demonstrate that considering the orientation of the viewing cameras relative to each point can aid data association and map culling in SLAM systems. A frequency count is maintained for each camera orientation bin and used to determine which points are likely to be visible from a given position for data association. In our work we maintain the two salient camera positions in each bin as opposed to a frequency count of all cameras.

3 Scene structure inference

In this section we propose a method that exploits both landmark position and the visibility and occlusion information from Section 2 to infer high-level semantic models describing the underlying scene structure. Without further constraint on the scene structure the problem is ill-posed as many explanations fit the data perfectly. One such explanation is that each landmark lies on an infinitesimal structure and each occluded camera is blocked by an infinitesimal occluding structure along the view ray to the landmark. We penalise such overfitting by evaluating the model evidence for competing explanations of the scene. Prankl *et al*. [15] use a method based on the Minimum Description Length criterion to fit planes to the scene. In our approach we evaluate the Bayesian model evidence directly following the method described in MacKay [10]. This approach permits the comparison of different classes of model, such as planes, bounded planes and non-planar point clusters. Complexity is penalised through the prior on model parameters and the sensitivity of the likelihood to small changes in the maximum likelihood model parameters. Additionally, by using a Bayesian method, we can avoid assigning points to models by marginalizing over all possible assignments.

3.1 Bayesian model evidence

The data available, D, consists of the landmark positions with associated visibility and occlusion information. Each hypothesis, $\mathcal{H}$ consists of a proposed model for the entire scene. The alternative hypotheses may be ranked by the evidence:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}) = \int P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})d\mathbf{w}P(\mathcal{H}), \tag{1}$$

where $\mathbf{w}$ is the vector of model parameters for the component models that constitute the hypothesis. The models we propose provide closed form first and second derivatives for the likelihood term given the model parameters, $P(D|\mathbf{w},\mathcal{H})$. This allows us to use Laplace's approximation for the marginalization in Equation 1 by using the MAP (maximum *a posteriori*) best fit for the model, $\mathbf{w}_{MP}$, and the Hessian. The model evidence used to rank the competing hypotheses is therefore taken to be:

$$P(D|\mathcal{H}) \approx P(D|\mathbf{w}_{MP},\mathcal{H})P(\mathbf{w}_{MP}|\mathcal{H})\ \det{}^{-1/2}(\mathbf{A}/2\pi). \tag{2}$$

where

$$\mathbf{A} = -\nabla\nabla \ln P(\mathbf{w}|D,\mathcal{H})|_{\mathbf{w}_{MP}} = -\nabla\nabla \ln P(D,\mathbf{w}|\mathcal{H})|_{\mathbf{w}_{MP}}. \tag{3}$$

Further details can be found in [10].

3.2 Mixture model assignment

The scene structure hypotheses are assembled from individual component models such as planes. The likelihood of a landmark given that it belongs to a particular model is a function of the point's 3D position in the model's coordinate frame. To determine the overall likelihood of a landmark with respect to the aggregate scene hypothesis it is necessary to assign landmarks to individual models. In the Bayesian model evidence framework we are able to marginalize over all possible assignments to yield an assignment agnostic score for the hypothesis. Avoiding hard assignment results in a cost function which does not have

discontinuities and is therefore easier to optimize. Additionally, points that are truly shared between models, such as those lying along the intersection of two planes, provide support for both models and not just one or the other. We consider the model evidence integrated over each assignment, $\mathbf{a}$, in the set of all possible assignments,

$$P(D|\mathcal{H}) = \int_{\mathbf{w}} \sum_{\mathbf{a}} P(D|\mathbf{w},\mathbf{a},\mathcal{H}) P(\mathbf{w}|\mathbf{a},\mathcal{H}) P(\mathbf{a}|\mathcal{H}) d\mathbf{w}. \tag{4}$$

The probability of a given assignment, $P(\mathbf{a}|\mathcal{H})$, is only conditioned on the hypothesis *but not its parameters* and so is determined only by the number of ways of assigning the N map points to M models, M^N. Also, we assume that in the absence of any data, the prior over the model parameters is independent of the assignment, $P(\mathbf{w}|\mathbf{a},\mathcal{H}) = P(\mathbf{w}|\mathcal{H})$.

With a view sphere bin size from Section 2 set such that the data can be considered i.i.d., as each data point, d_i, is independent of the assignment of the other data points,

$$\sum_{\mathbf{a}} P(D|\mathbf{w},\mathbf{a},\mathcal{H}) = \sum_{\mathbf{a}} \prod_i P(d_i|\mathbf{w},\mathbf{a},\mathcal{H}) \tag{5}$$

$$= \prod_i \sum_{a_i} P(d_i|\mathbf{w},a_i,\mathcal{H}), \tag{6}$$

where a_i is the assignment of the ith point. Therefore, using Laplace's approximation,

$$P(D|\mathcal{H}) = M^{-N} \prod_i \sum_{a_i} P(d_k|\mathbf{w}_{MP},a_i,\mathcal{H}) P(\mathbf{w}_M P|\mathcal{H}) \det{}^{-\frac{1}{2}} \left(\frac{\mathbf{A}}{2\pi}\right). \tag{7}$$

The log evidence is therefore

$$\ln P(D|\mathcal{H}) = \sum_i \ln \sum_{a_i} P(d_i|\mathbf{w}_{MP},a_i,\mathcal{H}) + \ln P(\mathbf{w}_{MP}|\mathcal{H}) - \tfrac{1}{2} \ln \left[\det\left(\frac{A}{2\pi}\right)\right] - N \ln M, \tag{8}$$

where

$$\mathbf{A} = -\nabla\nabla \ln P(\mathbf{w}|D,\mathcal{H}) = -\sum_i \nabla\nabla \ln \sum_{a_i} P(d_i|\mathbf{w},a_i,\mathcal{H})|_{\mathbf{w}_{MP}} - \nabla\nabla \ln P(\mathbf{w}|\mathcal{H})|_{\mathbf{w}_{MP}}. \tag{9}$$

3.3 Data likelihood

The data consists of the set of N landmark positions, $\mathbf{r} = \{r_1,\ldots,r_N\}$, together with the lists of visible, $\mathbf{v}_i = \{v_{i,1},\ldots,v_{i,K}\}$, and occluded, $\mathbf{o}_i = \{o_{i,1},\ldots,o_{i,K'}\}$, camera positions associated with each landmark, r_i. The likelihood of each landmark, d_i, is

$$P(d_i|a_i,\mathbf{w},\mathcal{H}) = P(\mathbf{v}_i,\mathbf{o}_i,r_i|a_i,\mathbf{w},\mathcal{H}) = P(r_i|a_i,\mathbf{w},\mathcal{H}) P(\mathbf{v}_i|r_i,\mathbf{w},\mathcal{H}) P(\mathbf{o}_i|r_i,\mathbf{w},\mathcal{H}). \tag{10}$$

The visibility and occlusion terms are taken to be independent of the assignment, and the position term is independent of the visibility. Omitting the $\mathcal{H}$ notation, the first term in Eq. 8 becomes

$$\ln \sum_{a_i} P(d_i|\mathbf{w}_{MP},a_i) = \ln \sum_{a_i} P(r_i|\mathbf{w}_{MP},a_i) + \sum_j \ln P(v_{i,j}|r_i,\mathbf{w}_{MP}) + \sum_j \ln P(o_{i,j}|r_i,\mathbf{w}_{MP}) \tag{11}$$

There are two explanations compatible with each camera in the visible list for a given landmark. The visible data point may have been generated either by successful feature matching in an unobstructed camera frame or by a false match in an obstructed frame. The visible likelihood is therefore

$$P(v_{i,j}|r_i,\mathbf{w}) = (1-\alpha)(1-P(b_{i,j}|r_i,\mathbf{w})) + \beta P(b_{i,j}|r_i,\mathbf{w}), \tag{12}$$

where $P(b_{i,j}|r_i,\mathbf{w})$ is the probability that the scene structure specified by $\mathcal{H}$ and $\mathbf{w}$ blocks the line from landmark r_i to camera centre $v_{i,j}$. The feature matching misdetection rate, α, and mismatch rate, β, account for false negatives and false positives in the visibility information.

The probability that the line from r_i to $v_{i,j}$ is blocked by a hypothesis consisting of M models is,

$$P(b_{i,j}|r_i,\mathbf{w}) = 1 - \prod_{m}^{M}(1-P(b_{i,j,m}|r_i,\mathbf{w}_m)), \tag{13}$$

where $P(b_{i,j,m}|r_i,\mathbf{w}_m)$ is the probability that r_i to $v_{i,j}$ is obstructed by model m. This probability is defined for each model class in Section 4.

The occluded data likelihood is likewise

$$P(o_{i,j}|r_i,\mathbf{w}) = \alpha(1-P(b_{i,j}|r_i,\mathbf{w})) + (1-\beta)P(b_{i,j}|r_i,\mathbf{w}). \tag{14}$$

3.4 Mixture model derivatives

The derivative of $P(r_i|\mathbf{w},a_i)$ with respect to the m^{th} model is weighted by the likelihood ratio of the landmark under model m to the sum over all models,

$$\frac{\partial}{\partial \mathbf{w}_m}\ln\sum_{a_i}P(r_i|\mathbf{w},a_i) = \left(\frac{1}{\sum_{a_i}P(r_i|\mathbf{w},a_i)}\right)\frac{\partial}{\partial \mathbf{w}_m}P(r_i|\mathbf{w}_m,a_i)|_{a_i=m}, \tag{15}$$

$$= \left(\frac{P(r_i|\mathbf{w},a_i)|_{a_i=m}}{\sum_{a_i}P(r_i|\mathbf{w},a_i)}\right)\frac{\partial}{\partial \mathbf{w}_m}\ln P(r_i|\mathbf{w}_m,a_i)|_{a_i=m}. \tag{16}$$

Each landmark is well explained by a few models at most, therefore the likelihood ratio is extremely small with respect to most models. A significant speed up is possible in implementation by culling contributions from models with likelihood ratios below a threshold so that the derivatives are only required from a handful of models for each data point. This threshold can be of the order of 10^{-6} whilst still providing a significant saving.

A similar speed up is possible for the visibility and occlusion terms. The derivative of $\ln P(v_{i,j}|r_i,\mathbf{w})$ is

$$\frac{\partial}{\partial \mathbf{w}_m}\ln P(v_{i,j}|r_i,\mathbf{w}) = \left(\frac{1-\alpha-\beta}{P(v_{i,j}|r_i,\mathbf{w})}\right)\frac{\partial}{\partial \mathbf{w}_m}P(b_i,j|r_i,\mathbf{w}) \tag{17}$$

where this model's contribution to $\frac{\partial}{\partial \mathbf{w}_m}P(b_{i,j}|r_i,\mathbf{w})$ is weighted by ratio of the probability that the view ray is unobstructed under model m to the probability that the ray is unobstructed by the entire hypothesis,

$$\frac{\partial}{\partial \mathbf{w}_m}P(b_{i,j}|r_i,\mathbf{w}) = \frac{(1-P(b_{i,j,m}|r_i,\mathbf{w}_m))}{(1-P(b_{i,j}|r_i,\mathbf{w}))}\frac{\partial}{\partial \mathbf{w}_m}P(b_{i,j,m}|r_i,\mathbf{w}). \tag{18}$$

The derivative of $\ln P(v_{i,j}|r_i,\mathbf{w})$ follows by the same method and an analytic closed form solution is also obtained for the Hessian.

4 Implementation

4.1 Populating the view spheres

The ultimate goal of the proposed method is to permit an online implementation running in parallel with a real-time structure from motion or visual SLAM system. An incremental update of the current scene structure hypothesis and the view spheres, running in parallel with the tracking and mapping system, could be exploited to improve data association by reasoning about visibility, occlusion and orientation of landmarks. In the current implementation the SFM point cloud is batch-processed to extract the view spheres. Klein and Murray's PTAM [9] is used to generate the initial landmarks and camera poses.

4.2 Model parameterization

The proposed framework permits heterogeneous models of various modality and dimensionality to be compared within a common framework. The model class and its parameters determine both the spatial probability distribution of map points generated by the model and the probability that the model obstructs camera j's view of landmark i. The simplest model proposed is an isotropic Gaussian distribution with a 3-vector, $\mathbf{p}$, for the position of the mean and a scalar, σ, for the variance. The likelihood of landmark r_i is simply

$$P(r_i|\mathbf{w}) = P(r_i|\mathbf{p},\sigma) = \frac{1}{(2\pi)^{3/2}\sigma^3}\exp\{\frac{(r_i-\mathbf{p})^\top(r_i-\mathbf{p})}{2\sigma^2}\} \tag{19}$$

The simplest version of this isotropic Gaussian model is transparent, so the probability of the model obstructing any view ray is zero.

The next most basic model is used to loosely model planes and extends the previous model to an anisotropic Gaussian by assigning a separate parameter, σ_z, to describe the variance in the z-axis of the model coordinate frame. This model requires three further parameters, two parameters to specify the direction of the z-axis relative to the world coordinate frame and a σ_z parameter that determines the variance in this direction. The rotation of the z-axis is given by the SO(3) parameterization, where the rotation parameters w_x and w_y are specified relative to the rotation of the plane, $\mathbf{R}$, at each iteration. Given $\Sigma = diag(\sigma_{xy},\sigma_{xy},\sigma_z)$, the likelihood of landmark r_i is

$$P(r_i|\mathbf{w}) = P(r_i|\mathbf{p},w_x,w_y,\sigma_{xy},\sigma_z) = \frac{1}{(2\pi)^{3/2}\sigma_{xy}^2\sigma_z}\exp\{\mathbf{q_i}^\top\Sigma^{-1}\mathbf{q_i}\}, \tag{20}$$

where $\mathbf{q_i} = \mathbf{R}^{-1}(r_i-\mathbf{p})$, the position of the landmark in plane coordinates. The extent in the xy-plane is governed by σ_{xy} enabling this model to describe both small planar patches and large planar structures in the scene. The limited extent of this plane model reduces interference with unrelated structure that plagues approaches which propose infinite planes.

Planes are then extended by adding boundaries and opacity. A convex bounded plane model exists when there are ≥ 3 boundary lines in the plane. The region inside the set of boundary lines forms a convex polygon and the distribution of points is uniform across this region in the xy-dimension. A sigmoid function is used to provide a smooth transition near the boundaries to permit continuous optimization of the boundary parameters. The normalization term is approximated by the area as the small error due to overlapping sigmoids at the

vertices is not significant. The distribution of points in the z-direction, normal to the plane, is a Gaussian parameterized by σ_z.

To consider opacity, the point of intersection with the plane and a view ray from a landmark to a camera determines the probability that the plane model obstructs this view. For the bounded plane model, the plane is transparent outside the boundaries and opaque inside, with a sigmoid function providing a continuous transition at the boundary. A smooth transition is also imposed on z-depth to transition from occluded, when a point lies behind the plane, to unoccluded in front of the plane.

The bounded plane is an obvious model to propose. The boundary parameters add complexity which is penalized when evaluating the Bayesian model evidence for the hypothesis. Alternative opaque planar models with fewer parameters have also been found to be useful in describing certain structure. For planes with ill-defined or hidden edges a semi-transparent plane model that exploits the σ_{xy} parameter, to define opacity by distance from the centre of the plane, requires fewer parameters than the bounded plane model. Whilst the semi-transparent model does not relate to a physical structure it is often useful as an intermediate stage in the optimization. Fully transparent, fully opaque and half-infinite planes are also proposed.

4.3 Optimization

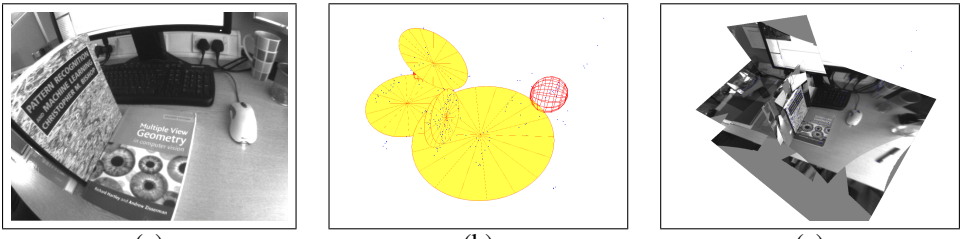
The scene structure is recovered incrementally by the addition, removal and replacement of models to generate new hypotheses at each iteration. The model evidence is evaluated at each stage and simulated annealing [4, 8] is employed to accept or reject the updated hypothesis. The choice of simulated annealing over the MCMC approach reflects that fact that we are only interested in best hypothesis and we do not require the probability distribution over the space of possible hypotheses. The proposal distribution is chosen to initialize the scene with simple isotropic Gaussian models before gradually upgrading to higher level structure.

Point models are proposed by uniformly sampling map points from the point cloud to initialize the point model position. Plane models are initialized from the visible map points in a given camera view using a RANSAC scheme. The method proposed by Myatt *et al*. [11] is used to draw the second and third points according to image distance from the first point to provide a good plane hypothesis.

The new hypothesis is optimized at each iteration, with the first and second order derivatives with respect to the model parameters available in analytic closed form. A trust region method with line searches is used to perform the optimization. This approach outperforms conjugate gradient and BFGS implementations due to the availability of the Hessian.

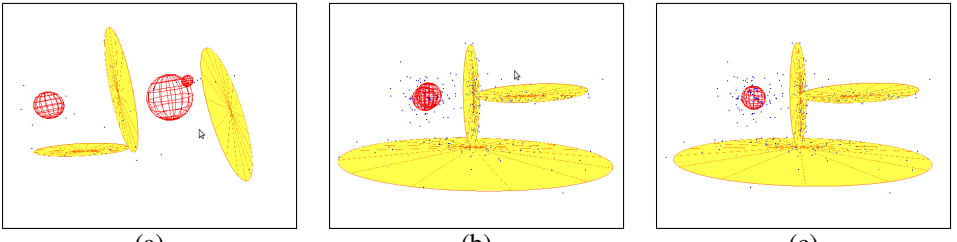
5 Results

Figure 2a is a typical frame taken from a video sequence of a desk scene with occlusion. Evidence of occlusion by the vertical book can be observed in the view sphere depicted in Figure 1. The scene structure hypothesis in Figure 2b is taken from an intermediate stage in the optimization. The large yellow discs represent the $3\sigma_{xy}$ radius in the plane of the semi-transparent plane models described in the previous section, whilst the red spheres are isotropic Gaussian models. Figure 2c has been generated by projecting the texture onto the planes from the camera view closest to the plane normal at each point.



(a) (b) (c)

Figure 2: Desk sequence with occlusion: (a) a typical video frame; (b) a snapshot of the scene structure hypothesis; (c) a novel view of the textured models.



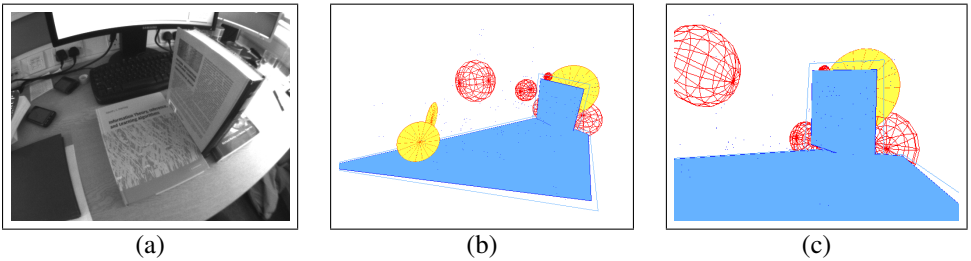
(a) (b) (c)

Figure 3: Synthetic experiment: (a) 10 landmarks per model; (b) 100; (c) ground truth.

Figure 3 shows the results of a synthetic experiment designed to investigate the impact of varying the number of landmarks found for each model in the scene. A synthetic scene is composed of two horizontal planes, a vertical plane and an isotropic Gaussian. In Figure 3a 10 landmarks were generated for each model in the scene. The vertical plane and isotropic Gaussian are successfully reconstructed along with a lower horizontal plane modelling a subset of the landmarks belonging to the true model. A false vertical plane and isotropic Gaussians are erroneously recovered to explain the remainder of the landmarks belonging to the horizontal places. As the number of landmarks generated per model is increased the reconstruction improves. At 100 landmarks per model the reconstruction is close to the ground truth.

In Figure 4 the set of proposed models is extended to include bounded plane models (depicted in blue). In Figure 4c the boundary of the book is accurately recovered by the visibility and occlusion information.

As the maximum number of camera positions stored in the viewsphere for each landmark is bounded the computational complexity of the view sphere representation grows linearly with map size. The complexity of the scene structure inference is $O(MN)$ where M is the number of models in the hypothesis and N is the number of landmarks in the scene. The cur-



(a) (b) (c)

Figure 4: Scene structure hypothesis: (a) video frame; (b) & (c) bounded plane model.

rent non-optimized batch implementation takes several seconds to optimize each proposed hypothesis for a typical desk scene. To permit a comparison with methods such as [12] an online implementation is needed in future work. Complexity may be reduced by exploiting the local influence of individual models in an hypothesis.

6 Conclusions

Visibility and occlusion information provides useful evidence for the presence of otherwise undetected scene structure and correctly positions object boundaries in the absence of texture. Bayesian model evidence provides an effective measure of the quality of a scene structure hypothesis. The framework proposed can be extended to incorporate higher order models including volumes and concepts of parallelism, orthogonality and repeated structure.

The two main contributions of this paper, a method for the efficient extraction of visibility and occlusion information with low redundancy and a framework for generating high-level models compatible with this information, have been shown to provide an effective description of scene structure from sparse primitives, with the potential to scale to a real-time implementation in future work.

References

[1] S. Carlsson. Duality of reconstruction and positioning from projective views. In *Proc. IEEE Workshop on Representation of Visual Scenes*, 1995.

[2] J.M. Carranza and A. Calway. Efficiently increasing map density in visual slam using planar features with adaptive measurement. In *Proc. 20th British Machine Vision Conference*, 2009.

[3] J.M. Carranza and A. Calway. Unifying planar and point mapping in monocular slam. In *Proc. 21st British Machine Vision Conference*, 2010.

[4] V. Cerny. A thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of optimization theory and applications*, 45(1):41–51, 1985. ISSN 0022-3239.

[5] A.R. Dick, P.H.S. Torr, and R. Cipolla. Modelling and interpretation of architecture from several images. *International Journal of Computer Vision*, 2004.

[6] A.P. Gee, D. Chekhlov, A. Calway, and W. Mayol-Cuevas. Discovering higher level structure in visual slam. *IEEE Transactions on Robotics*, 2008.

[7] R. Hartley and G. Debunne. Dualizing scene reconstruction algorithms. In *Proc. European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, 1998.

[8] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983. ISSN 1095-9203.

[9] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *Proc. 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007.

[10] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003. ISBN 0521642981.

[11] D.R. Myatt, P.H.S. Torr, S.J. Nasuto, J.M. Bishop, and R. Craddock. NAPSAC: High noise, high dimensional robust estimation - it's in the bag. In *Proc. 13th British Machine Vision Conference*, 2002.

[12] R.A. Newcombe and A.J. Davison. Live dense reconstruction with a single moving camera. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[13] Q. Pan, G. Reitmayr, and T. Drummond. Proforma: Probabilistic feature-based on-line rapid model acquisition. In *Proc. 20th British Machine Vision Conference*, 2009.

[14] K. Pirker, M. Ruther, and H. Bischof. Histogram of oriented cameras - a new descriptor for visual slam in dynamic environments. In *Proc. 21st British Machine Vision Conference*, 2010.

[15] J. Prankl, M. Zillich, B. Leibe, and M. Vincze. Incremental model selection for detection and tracking of planar surfaces. In *Proc. 21st British Machine Vision Conference*, 2010.

[16] S. Wangsiripitak and D.W. Murray. Reducing mismatching under time-pressure by reasoning about visibility and occlusion. In *Proc. 21st British Machine Vision Conference*, 2010.