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

We propose a method for extracting high-level scene structure from the type of data provided by 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.

A visual SLAM system operates by matching landmarks in a map into images from multiple viewpoints. The success or failure of the matching contains some information about the visibility or occlusion of that landmark from those views. For a system operating at frame-rate the occlusion information provided by each camera pose is highly correlated with the previous frame and much of the occlusion information is redundant. Redundancy is illustrated in Figure 1. The red camera marked with $\star$ indicates that the landmark is occluded by unmodelled structure. This information is redundant given the presence of the closer camera. The principle applies in reverse for cameras which successfully observe points.

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.

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. 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.

We propose a method that exploits this visibility and occlusion information 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. We avoid over-fitting by evaluating the Bayesian model evidence for competing explanations of the scene following the method described in MacKay [1]. 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.

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:

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

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:

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

where

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

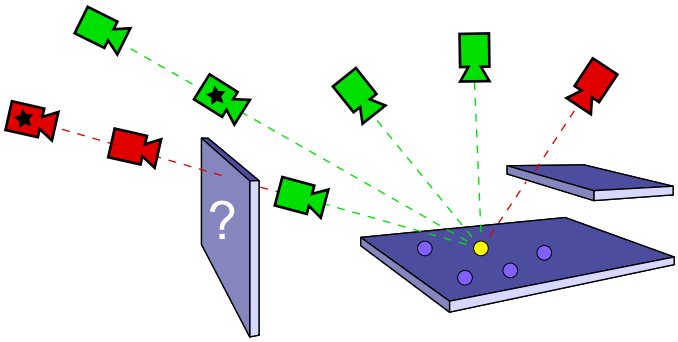

Figure 1: Visibility and occlusion: Successful (green) and unsuccessful (red) observation of a landmark (yellow) provides evidence for unmodeled structure causing occlusion (plane marked with '?'). Cameras marked with $\star$ provide redundant information about occlusion.

The scene structure hypotheses are assembled from individual component models such as planes. 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,

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

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, $P(\mathbf{w}|\mathbf{a},\mathcal{H}) = P(\mathbf{w}|\mathcal{H})$.

Each data point, d_i, can be considered independent of the assignment of the other data points,

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

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

$$P(D|\mathcal{H}) = M^{-N} \prod_i \sum_{a_i} P(d_k|\mathbf{w}_{MP},a_i,\mathcal{H})P(\mathbf{w}_M P|\mathcal{H})\det{}^{-\frac{1}{2}}\left(\frac{\mathbf{A}}{2\pi}\right). \qquad (6)$$

The implementation of this method is described in the paper, including details of the model parameterization and non-linear optimization. Our conclusion is that 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.

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