Layer Extraction with a Bayesian model of shapes

P. H. S. Torr[1], A. R. Dick[2], and R. Cipolla[2]

[1] Microsoft Research, 1 Guildhall St, Cambridge CB2 3NH, UK
philtorr@microsoft.com
[2] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
{ard28,cipolla}@eng.cam.ac.uk

Abstract. This paper describes an automatic 3D surface modelling system that extracts dense 3D surfaces from uncalibrated video sequences. In order to extract this 3D model the scene is represented as a collection of layers and a new method for layer extraction is described. The new segmentation method differs from previous methods in that it uses a specific prior model for layer shape. A probabilistic hierarchical model of layer shape is constructed, which assigns a density function to the shape and spatial relationships between layers. This allows accurate and efficient algorithms to be used when finding the best segmentation. Here this framework is applied to architectural scenes, in which layers commonly correspond to windows or doors and hence belong to a tightly constrained family of shapes.

Keywords: Structure from motion, Grouping and segmentation.

1 Introduction

The aim of this work is to obtain dense 3D structure *and* texture maps from an image sequence, the camera matrices (calibration and location) having been recovered using previously developed methods [3, 4, 12, 15]. The computed structure can then be used as the basis for building 3D graphical models. This representation can be used as a basis for compression, new view rendering, and video editing. A typical example sequence is shown in Figure 1 and the computed model in Figure 9.

Although extracting scene structure using stereo has been actively researched, the accurate recovery of the depth for each pixel remains only partially solved. For instance, one approach to the dense stereo problem is the voxel based approach [14] in which the scene volume is first discretized into voxels, and then a space carving scheme applied to find the voxels that lie on the surfaces of the objects in the scene. The disadvantage of the voxel carving method is that the surfaces produced from homogeneous regions are “fattened” out to a shape known as the Photo Hull [14]. Rather than generate voxels in 3D some algorithms operate in the image by testing different disparities for each pixel e.g.

Koch *et al* [9]. The problem with these approaches are that they do not treat all the images equally and work well only for small baselines.

Generally dense stereo algorithms work well in highly textured regions, but perform poorly around occlusion boundaries and in untextured regions. This is because there is simply not enough information in these untextured regions to recover the shape. In this paper we propose a general framework for overcoming this by the utilization of prior knowledge.

A vehicle for encoding this prior knowledge is the decomposition of the image into layers [1, 2, 8, 17–19]. Each layer corresponds to a surface in the scene, hence the decomposition of the scene into layers acknowledges the conditional dependence of the depths for adjacent pixels in an image. Detecting the different surfaces (layers) within the scene offers a compact and physically likely representation for the image sequence. The main problem is that such a decomposition is difficult to achieve in general. This is because the parametrization of the layer itself is problematic. For each layer the parametrization is composed of three parts: (a) the parameteric form of the 3D surface giving rise to the layer, (b) its spatial extent within the image and (c) its texture map. Generally it is easy to construct the former e.g. in [1, 8, 18] it is assumed that the surfaces are planar, in [2, 17] the surfaces are encoded by a plane together with a per pixel parallax, in [19] only smoothness of depth is assumed. The latter two however are more difficult to parametrize. One approach is to ignore the spatial cohesion altogether and simply model the whole image as a mixture model of the layers [1, 8, 19]. Whilst this simplifies the problem of estimating the layers affording the use of iterative algorithms like EM, it is not a realistic model of layer generation e.g. a homogeneous region which contains little depth or motion information could be broken up in any way and assigned to different layers with no increase in the mixture model's likelihood.

A now classical method for modelling the spatial dependence of layer memberships of adjacent pixels is by use of Markov Random Fields (MRFs) [7]. There are several disadvantages with this approach: first is that using an MRF model leads to very difficult optimization problems that are notoriously slow to converge. Second, sampling from an MRF distribution does not produce things that look like images of the real world, which might lead one to think that using this as a prior is a bad idea. Third, the MRF is pixel based which can lead to artefacts. The MRF only implicitly defines the prior probability distribution in that the normalization factor cannot be readily computed. What would be preferable would be an explicit prior for the segmentation, which would allow more direct minimization of the error function, for instance by gradient descent.

Within this paper a prior for the shape of the layers is constructed and illustrated for architectural scenes. Architectural scenes are particularly amenable to the construction of priors, as layers will typically correspond to such things as windows or doors which for which an informative prior distribution can be constructed (e.g. they are often planar with regular outline). Although architectural scenes are chosen to illustrate the basic principles the method proposed is representative of *a general approach* to segmentation. Taking inspiration from [6], rather than using an implicit model for the prior probability of a segmentation an explicit model is defined and used. A solution to the final problem (c), that of finding the texture map, is also found by considering the texture as a set of hidden variables.

The layout of the paper is as follows. The parameters used to represent the shape and texture of a scene are defined in Section 2. A posterior probability measure is also introduced here for estimating the optimal parameter values for a scene from a set of images and prior information. As shape is now represented parametrically, layer extraction becomes a problem of *model selection*, i.e. determining the number and type of these parameters required to model the scene. In Section 3 a method is developed for choosing automatically which model is most appropriate for the current scene, based on goodness of fit to the images and the idea of model simplicity. Section 4 then deals with the details of implementing this method. In Section 5 it is demonstrated that this technique can decide which individual shape model is appropriate for each layer, which overall model best fits the collection of layers in the scene, and how many layers are present in the scene, given a coarse initialisation. Concluding remarks are given in Section 6.

2 Problem formulation

A scene is modelled as a collection of layers. A scene model has a set of parameters represented by a vector $\boldsymbol{\theta}$, which can be decomposed into shape parameters $\boldsymbol{\theta}_S$ and texture parameters $\boldsymbol{\theta}_T$ such that $\boldsymbol{\theta} = \boldsymbol{\theta}_S \bigcup \boldsymbol{\theta}_T$. Each layer is defined as a deformable template in three space; the shape parameters $\boldsymbol{\theta}_S$ comprise the location and orientation of each template together with the boundary (a variable number of parameters for each layer depending on which model $\mathcal{M}$ is selected). A grid is defined on the bounded surface of each layer, and each point on this grid is assigned an intensity value forming a texture map on the 3D layer. The intensity at each grid point is a variable of $\boldsymbol{\theta}_T$. The projection matrix to a given image, a noise process, and $\boldsymbol{\theta}$ provide a complete generative model for that image. Each point in the model can be projected into the image and the projected intensity compared with that observed, from which the likelihood of the model can be computed. If priors are assigned to the parameters then the posterior likelihood can be computed.

Within this paper a dominant plane is assumed to fill most of the scene (such as a wall of the building), with several offset objects (such as windows, doors and pillars). An example of such a scene is given in Figure 1. The background plane $\mathcal{L}_0$ is modelled as the plane $z = 0$ with infinite extent (thus having no shape parameters). The other layers $\mathcal{L}_1 \ldots \mathcal{L}_m$ are modelled as deformable templates as now described.

2.1 The shape parameters

At present there are four types of layer model $\mathcal{M}$ available, which allow the modelling of a wide variety of architectural scenes. These are $\mathcal{M}_1$ a rectangle (6

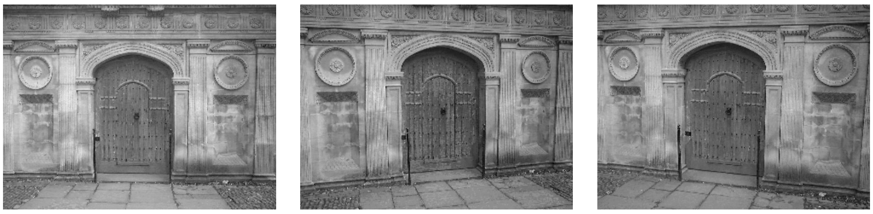


Fig. 1. *Three images of the type of scene considered. A gateway and two indentations are offset from the background plane of the wall.*

parameters), $\mathcal{M}_2$ an arch (7 parameters), $\mathcal{M}_3$ a rectangle with bevelled (sloped) edges (7 parameters), and $\mathcal{M}_4$ an arch with bevelled edges (8 parameters). The 8 parameter model $\mathcal{M}_4$ has position coordinates (x, y), scale parameters a and b, orientation ω, an arch height c, depth from the background plane d and bevel width r. The arch in $\mathcal{M}_2$ and $\mathcal{M}_4$ is completely specified by c as it is modelled using a semi-ellipse. The other layer models are constrained versions of this model, as shown in Figure 2.

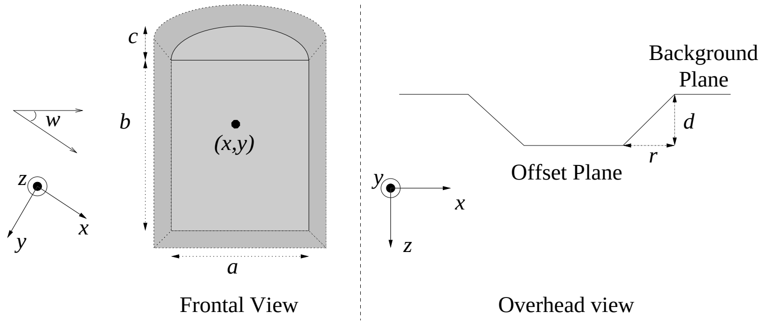


Fig. 2. *Top and cross-sectional views of the most general shape primitive. The other primitives are special cases of this one—the non-bevelled arch has $r = 0$, the bevelled rectangle has $c = 0$ and the non-bevelled rectangle has $r = 0, c = 0$. The coordinate axes shown for each view are translated versions of the 3D world coordinate system.*

Layers in architectural scenes are highly constrained not only in their individual shape, but also in their spatial relationship to each other. Hence a single parameter can often be used to represent a feature common to several primitives, such as the common y position of layers belonging to a single row. These global parameters are known as *hyperparameters* [5], as the entities that they model are themselves parameters. The introduction of hyperparameters makes the model hierarchical as illustrated in the directed acyclic graph (DAG) Figure 3. The hyperparameters defined in Table 1 are later used to represent our belief that primitives occur in rows, but there are many other possibilities.

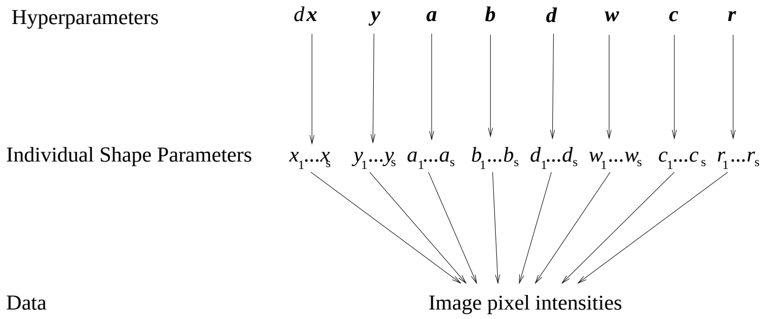


Fig. 3. *The hierarchical shape model. Hyperparameters model functions of the individual shape parameters. The camera projection matrices, and lighting conditions, could also be modelled as hyperparameters based on the data and shape parameters, but in this paper they are given as prior information.*

To sum up, architectural scenes containing a background layer $\mathcal{L}_0$, together with a set of offset layers $\mathcal{L}_i$, $i = 1 \dots m$ are to be modelled. Each offset layer has an associated model $\mathcal{M}_j$, $j = 1 \dots 4$. The individual shape parameters and the hyperparameters together define the shape of the model, and can be represented as a parameter vector $\boldsymbol{\theta}_S$; next the texture paramters $\boldsymbol{\theta}_T$ are defined.

2.2 Texture Parameters

The set of layers defined above define a surface. Next this surface is discretized and a two dimensional coordinate system defined on it. At each point $\mathbf{X}$ on this surface an unknown brightness parameter $i(\mathbf{X})$ (between 0 and 255) is defined. These brightness parameters form the texture parameter vector $\boldsymbol{\theta}_T$.

2.3 Evaluating the Likelihood

The shape parameters (number of layers and their associated parameters) and the texture parameters give the total parameter vector $\boldsymbol{\theta}$. In order to estimate this its posterior probability must be maximized:

$$p(\boldsymbol{\theta}|\mathbf{DI}) = p(\mathbf{D}|\boldsymbol{\theta}\mathbf{I})p(\boldsymbol{\theta}|\mathbf{I}) \qquad (1)$$

where $\mathbf{I}$ is the prior information (such as the camera matrices etc.) and $\mathbf{D}$ is the set of input images. This is a product of the likelihood and prior. To perform the optimization gradient descent is used. This would prove prohibitive if all

Table 1. *Example set of hyperparameters. Knowledge about overall scene structure can be imposed by assigning a probability distribution to each hyperparameter.*

$d\mathbf{x}$	The spacing of x-axis position of the primitives
$\mathbf{y}$	The y-axis position of the primitives
$\mathbf{a}$	The horizontal scale of the primitives
$\mathbf{b}$	The vertical scale of the primitives
$\mathbf{d}$	The depth of the primitives
$\boldsymbol{\omega}$	The orientation of the primitives
$\mathbf{c}$	The arch height of the primitives
$\mathbf{r}$	The bevel width of the primitives

the paramaters had to be searched simultaneously. Fortunately the task can be decomposed into several easier optimizations: first the shape parameters of each layer can be optimized independently, second only the shape parameters need to be optimized explicitly. It is now shown how to estimate the optimal set of texture parameters given these shape parameters.

Given the shape parameters and projection matrices, it is now assumed that the projected intensity of $\mathbf{X}$ is observed with noise $i(\mathbf{X}) + \epsilon$, where ϵ has a Gaussian distribution mean zero and standard deviation σ_ϵ. The parameter $i(\mathbf{X})$ can then be found such that it minimizes the sum of squares $\min_{i(\mathbf{X})} \sum_{j=1}^{j=m} \left(i(\mathbf{x}^j) - i(\mathbf{X})\right)^2$ where $i(\mathbf{x}^j)$ is the intensity at $\mathbf{x}^j$, and $\mathbf{x}^j$ is the projection of $\mathbf{X}$ into the jth image. The likelihood for a given value of $i(\mathbf{X})$ is

$$p(\mathbf{D}|i(\mathbf{X})) = \prod_j \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp -\frac{1}{2}\left(\frac{i(\mathbf{x}^j) - i(\mathbf{X})}{\sigma_\epsilon}\right)^2 \tag{2}$$

Using Equation (2) under the assumption that the errors ϵ in all the pixels are independent, the likelihood over all pixels can then be written:

$$p(\mathbf{D}|\boldsymbol{\theta}_T\boldsymbol{\theta}_S\mathbf{I}) = \prod_i \prod_j \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp -\frac{1}{2}\left(\frac{i(\mathbf{x}_i^j) - i(\mathbf{X}_i)}{\sigma_\epsilon}\right)^2 \tag{3}$$

where $\mathbf{x}_i^j$ is the projection of the ith scene point into the jth image. This summation is over all the discretized scene points (lying on the surfaces of the layers) $\mathbf{X}_i$.

2.4 Evaluating the priors

Prior knowledge of parameter values is encoded in the prior probability term of Equation (1), $p(\boldsymbol{\theta}|\mathbf{I}) = p(\boldsymbol{\theta}_S\boldsymbol{\theta}_T|\mathbf{I}) = p(\boldsymbol{\theta}_S|\mathbf{I})$, as the value of the texture parameters is determined by the shape parameters and the images. The shape parameter vector $\boldsymbol{\theta}_S = (\alpha, \beta)$ contains both individual shape parameters α and hyperparameters β. Hence the prior probability $p(\boldsymbol{\theta}_S|I) = p(\alpha\beta|I) = p(\alpha|\beta I)p(\beta|I)$.

Table 2. *Hyperpriors encoding a row of identical primitives. $U[a,b]$ is the uniform distribution over the interval $[a,b]$. $N(a,b)$ is the normal distribution with mean a and standard deviation b. A column of primitives is similarly constrained, by imposing a hyperprior on* $\mathbf{x}$ *and* $d\mathbf{y}$. *The model typically has a spatial extent of [-0.5, 0.5] in the x and y axes of the world coordinate system; hence a standard deviation of 0.005 corresponds to 2 or 3 pixels in a typical image of the scene.*

β	$p(\beta)$	$p(\alpha\|\beta)$
$d\mathbf{x}$	$U[0.2, 0.4]$	$N(d\mathbf{x}, 0.005)$
$\mathbf{y}$	$U[-0.4, 0.4]$	$N(\mathbf{y}, 0.005)$
$\mathbf{a}$	$U[0.1, 0.2]$	$N(\mathbf{a}, 0.005)$
$\mathbf{b}$	$U[0.1, 0.2]$	$N(\mathbf{b}, 0.005)$
$\mathbf{d}$	$U[-0.1, 0.1]$	$N(\mathbf{d}, 0.005)$
$\boldsymbol{\omega}$	$N(0, \pi/12]$	$N(\boldsymbol{\omega}, \pi/12)$
$\mathbf{c}$	$U[0.01, 0.2]$	$N(\mathbf{c}, 0.005)$
$\mathbf{r}$	$U[0.01, 0.1]$	$N(\mathbf{r}, 0.005)$

The term $p(\beta|I)$, known as a *hyperprior*, expresses a belief in the overall structure of the scene, while $p(\alpha|\beta I)$ determines how individual shapes in the scene are expected to vary within the overall structure. To express complete prior ignorance about the scene structure, each prior probability may be assigned a uniform distribution bounded by the range of the cameras' fields of view. The correct distribution for each hyperparameter should ideally be learnt automatically from previous data sets; however at present they are manually initialised. An example of a set of hyperpriors for a row of identical shapes is given in Table 2. Samples from this distribution are given in Figure 4.

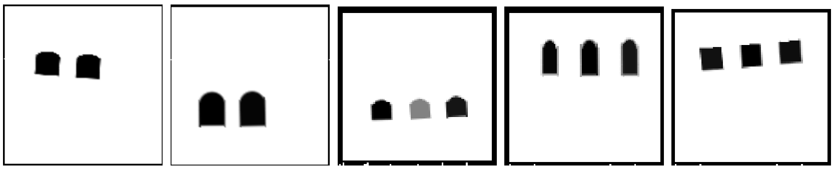


Fig. 4. *Samples drawn from the hyperprior distribution for a row of identical primitives given in table 2, using two and three primitives. The intensity at each point is proportional to the depth offset from the background layer.*

3 Model selection

In Section 2 a set of parameters was defined, and the posterior probability (Equation (1)) was introduced as a means of estimating the optimal parameter values for a given model. However a more fundamental problem remains: how to decide which model (i.e. which set of parameters) best represents a scene? This is the problem of *model selection*, described in this section.

The goal of model selection is to choose the most probable of a finite set of models $\mathbf{M}_j, j = 1..n$, given data $\mathbf{D}$ and prior information $\mathbf{I}$. Using Bayes rule the probability of each model can be expressed as

$$p(\mathbf{M}_j|\mathbf{DI}) = \frac{p(\mathbf{D}|\mathbf{M}_j\mathbf{I})p(\mathbf{M}_j|\mathbf{I})}{p(\mathbf{D}|\mathbf{I})}. \tag{4}$$

The denominator $p(\mathbf{D}|\mathbf{I})$ is constant for all models and hence is used only as a normalisation constant to ensure that $\sum_{j=1}^{n} p(\mathbf{M}_j|\mathbf{DI}) = 1$. The prior probability $p(\mathbf{M}_j|\mathbf{I})$ can be used to encode any prior preference one has for each model. In the absence of any such prejudice this is uniform, and model selection depends primarily on the *evidence*

$$p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) = \int p(\mathbf{D}|\boldsymbol{\theta}_j\mathbf{M}_j\mathbf{I})p(\boldsymbol{\theta}_j|\mathbf{M}_j\mathbf{I})d\boldsymbol{\theta}_j \tag{5}$$

where $\boldsymbol{\theta}_j$ is the set of parameters belonging to model $\mathbf{M}_j$.

For this problem, the data $\mathbf{D}$ is simply a set of images of the scene. The prior information $\mathbf{I}$ is the projection matrix for each camera, and a noise model for projection into each image (Section 2.3). The parameter vector $\boldsymbol{\theta}_j$ contains shape and texture parameters, as described in Section 2. Considering these separately, the evidence becomes

$$\begin{aligned} p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) &= \int\int p(\mathbf{D}|\alpha_j\beta_j\boldsymbol{\theta}_{Tj}\mathbf{M}_j\mathbf{I})p(\alpha_j\beta_j\boldsymbol{\theta}_{Tj}|\mathbf{M}_j\mathbf{I})d\alpha_j d\beta_j && (6)\\ &= \int\int p(\mathbf{D}|\alpha_j\mathbf{M}_j\mathbf{I})p(\alpha_j|\beta_j\mathbf{M}_j\mathbf{I})p(\beta_j|\mathbf{M}_j\mathbf{I})d\alpha_j d\beta_j. && (7)\end{aligned}$$

The β_j term is dropped from the first factor of this equation, the *likelihood* of the data, as the probability of the data $\mathbf{D}$ is dependent only on the individual shape parameters α_j. The texture parameters are not considered here as they are completely determined by the shape parameters and the images.

3.1 Evaluation of the evidence

It is impractical to perform the integration of Equation (6) for any but the simplest models. However previous work (e.g. [13, 16, 10]) has shown that an approximation to the evidence is sufficient for model selection. Five possible approximations are briefly described here. Consider first the inner integral,

$$\int p(\mathbf{D}|\alpha_j\mathbf{M}_j\mathbf{I})p(\alpha_j|\beta_j\mathbf{M}_j\mathbf{I})p(\beta_j|\mathbf{M}_j\mathbf{I})d\alpha_j. \tag{8}$$

In any useful inference problem the likelihood fuction obtained from the data will be much more informative than the prior probabilities—if not, there is little to be gained by using the data. Hence the likelihood function has a sharp peak (relative to the prior distribution) around its maximum value, as shown in Figure 5(a). The entire integral is therefore well approximated by integrating a neighbourhood of the maximum likelihood estimate of α_j. For convenience the log likelihood, $\log L(\alpha_j)$, which is a monotonic function of the likelihood and hence is maximised at the same value, is considered rather than the likelihood itself. A second order approximation of $\log L(\alpha_j)$ about its mode α_{jML} is

$$\log L(\alpha_j) \cong \log L(\alpha_{jML}) + \frac{1}{2}(\alpha_j - \alpha_{jML})^T \mathbf{H}(\alpha_{jML})(\alpha_j - \alpha_{jML}), \qquad (9)$$

where $\mathbf{H}(\alpha_{jML})$ is the hessian of $\log L$ evaluated at its mode. This corresponds to approximating the likelihood in this region by a multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma}_\alpha = -\mathbf{H}^{-1}(\alpha_{jML})$:

$$L(\alpha_j) \cong L(\alpha_{jML}) \exp\left[-\frac{1}{2}(\alpha_j - \alpha_{jML})^T \boldsymbol{\Sigma}_\alpha^{-1}(\alpha_j - \alpha_{jML})\right] \qquad (10)$$

Assuming that $p(\alpha_j|\beta_j\mathbf{M}_j\mathbf{I}) \cong p(\alpha_{jML}|\beta_j\mathbf{M}_j\mathbf{I})$ in this region, the integrand of Equation (8) reduces to

$$\int \exp\left[-\frac{1}{2}(\alpha_j - \alpha_{jML})^T \boldsymbol{\Sigma}_\alpha^{-1}(\alpha_j - \alpha_{jML})\right] d\alpha_j = (2\pi)^{k_\alpha/2}\sqrt{\det(\boldsymbol{\Sigma}_\alpha)} \qquad (11)$$

where k_α is the number of shape parameters, as a Gaussian must integrate to 1. Hence the integral of Equation (8) is approximately

$$L(\alpha_{jML})(2\pi)^{k_\alpha/2}\sqrt{\det \boldsymbol{\Sigma}_\alpha} p(\alpha_{jML}|\beta_j\mathbf{M}_j\mathbf{I})p(\beta_j|\mathbf{M}_j\mathbf{I}) \qquad (12)$$

and the evidence is approximated as

$$p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) \cong L(\alpha_{jML})(2\pi)^{k_\alpha/2}\sqrt{\det \boldsymbol{\Sigma}_\alpha}\int p(\alpha_{jML}|\beta_j\mathbf{M}_j\mathbf{I})p(\beta_j|\mathbf{M}_j\mathbf{I})d\beta_j \qquad (13)$$

The remaining integral can be similarly approximated; now the parameter values α_{jML} are the "data" being used to estimate the hyperparameters β_j. The final expression for the evidence is therefore

$$p(\mathbf{D}|\mathbf{M}_j\mathbf{I}) \cong L(\alpha_{jML})p(\alpha_{jML}|\beta_{jML}\mathbf{M}_j\mathbf{I})p(\beta_{jML}|\mathbf{M}_j\mathbf{I})(2\pi)^{k/2}\sqrt{\det \boldsymbol{\Sigma}_\alpha \det \boldsymbol{\Sigma}_\beta}. \qquad (14)$$

where k is the total number of parameters, $\boldsymbol{\Sigma}_\beta$ is the covariance matrix of the hyperparameters with respect to the shape parameters and β_{jML} is the set of hyperparameters which maximise $p(\alpha_{jML}|\beta_j\mathbf{M}_j\mathbf{I})$. The terms to the right of the likelihood approximate the fraction of the volume of prior probability space enclosed by the maximum likelihood peak, and are known collectively as an *Occam factor* [10, 16].

3.2 Occam factors

An Occam factor encodes the idea of Occam's Razor for model selection: a simpler model (usually one with fewer parameters) should be preferred to a more complex one, unless the more complex one explains or fits the data significantly better. Information theoretic techniques such as Minimum Description Length (MDL) encoding [1] enforce this preference by penalising models according to the information required to encode them. The Bayesian approach to model selection naturally incorporates an identical penalty in the evaluation of the evidence as the product of a likelihood and an Occam factor. As extra parameters are introduced into the model, the fraction of the volume of parameter space occupied by the peak surrounding the maximum likelihood estimate inevitably decreases, as illustrated in Figure 5 for the simple case of 1 and 2 parameter models. Because the volume of prior probability over the entire parameter space must always be 1, this decrease in occupied volume translates to a decrease in prior probability, assuming that prior probability is reasonably uniform over the parameter space.

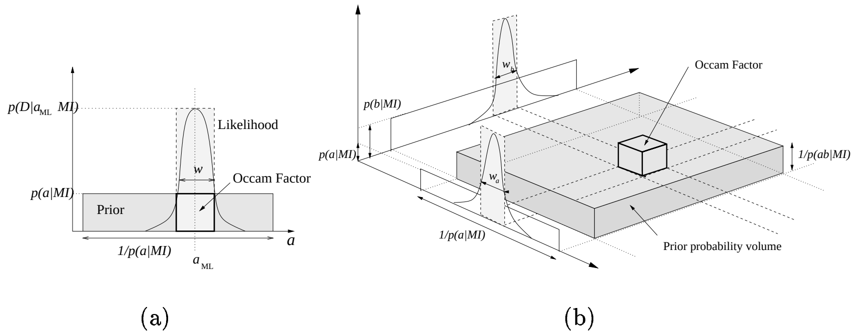


Fig. 5. *Occam factor for (a) one and (b) two independent parameters with uniform independent priors. For a single variable, the likelihood is approximated as* $p(\mathbf{D}|a_{ML}\mathbf{MI})$ *and the Occam factor is* $w/p(a|\mathbf{MI})$*. In the case of two variables the likelihood is* $p(\mathbf{D}|a_{ML}b_{ML}\mathbf{MI})$ *and the Occam factor is* $w_a w_b/p(ab|\mathbf{MI})$*. In each case the Occam factor approximates the fraction of the volume of prior probability (the shaded volume) occupied by the maximum likelihood probability peak.*

3.3 Other approximations to the evidence

The Occam factor is one of many possible model selection criteria. By disregarding it completely, evidence evaluation reduces to a maximum likelihood (ML) estimation. This includes no preference for model parsimony and hence will always select the best fitting model regardless of its complexity. If the prior terms $p(\alpha_{jML}|\beta_{jML}\mathbf{M}_j\mathbf{I})$ and $p(\beta_{jML}|\mathbf{M}_j\mathbf{I})$ are known to be almost uniform, Schwarz[13] suggests approximating them with diffuse normal distributions, and approximating $\sqrt{\det \boldsymbol{\Sigma}_\alpha \det \boldsymbol{\Sigma}_\beta}$ by $N^{\frac{-k}{2}}$, where N is the number of observations and k is the number of parameters in the model. This forms the Bayesian Information Criterion (BIC) measure of evidence

$$\log(p(\mathbf{D}|\mathbf{M}_j\mathbf{I})) \cong \log L(\alpha_{jML}) - \frac{k}{2}\log N. \tag{15}$$

A non Bayesian penalty term, the AIC has the form

$$\log(p(\mathbf{D}|\mathbf{M}_j\mathbf{I})) \cong \log L(\alpha_{jML}) - 2k \tag{16}$$

and hence penalises models according to the number of parameters they include. Finally if the posterior distribution is very peaked, the MAP estimate of each model may be the same order of magnitude as the evidence, in which case one would expect it to perform just as well for model selection. Each of these criteria is compared in section 5.

4 Implementation issues

4.1 Initialisation

The purpose of the initialisation stage is to provide a rough estimate of the number of planes to be modelled, and their position, scale, depth and orientation. First, each image is warped by a transformation $\mathbf{A}_j$ so that the layer $\mathcal{L}_0$ is aligned. Approximate projection matrices are found by estimating each camera pose from $\mathbf{A}_j$, as in [15]. A dense parallax field is obtained by applying a wavelet transform to each warped image, and performing multiresolution matching in the phase domain [11]. The correspondences obtained from each pair of images are fused robustly to obtain depth estimates for each point, from which initial layer estimates can be hypothesised [4].

Initial parameter estimates are obtained by fitting the simplest model, a 6 parameter rectangle, to each region (see Figure 6). The centre of the rectangle is positioned at the centroid of the region. The horizontal and vertical scales are set to the average distance of each of the extrema of the region from the centroid, the depth is given by the depth of the centroid and the orientation is assumed to be vertical (i.e. $\omega = 0$). The projection matrices generated by this system have a typical reprojection error of order 1 pixel.

4.2 Search for the maximum likelihood parameters

A multiresolution gradient descent search is used to locate the maximum likelihood parameters α_{jML} for each possible shape model. The image is recursively convolved with a Gaussian filter and downsampled by a factor of 2 horizontally and vertically to obtain a multiresolution representation. The search is initialised

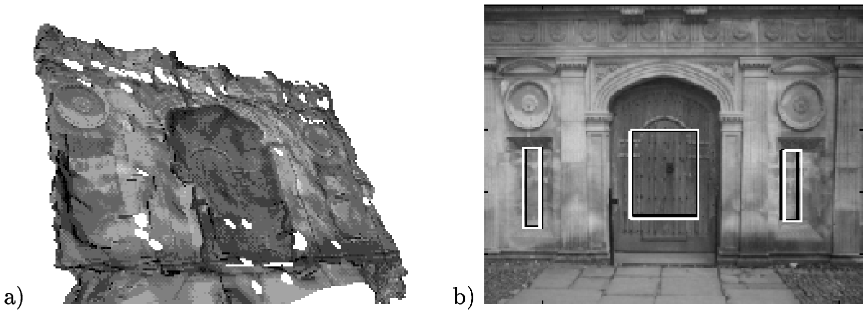


Fig. 6. *(a) A poor initial reconstruction (compare with Figure 9) based only on stereo between two images, and (b) initial layer estimates based on three images. The layers are bounded by the black lines and meet the zero layer at the white lines. Each major offset layer is detected, but their shape and size are estimated poorly.*

at the coarsest level, and the estimate found is used to seed the search at a finer level. At each level the model is sampled more densely to maintain a constant sample rate of approximately one point per image pixel. Experience shows that two or three levels of resolution are sufficient, and the search typically converges in less than 100 iterations.

5 Results

5.1 Model selection for the shape parameters

Initially the model selection algorithm is assessed by trying to identify the correct shape for a single layer. Starting from the parameters found during initialisation (section 4), the gradient descent method described in Section 4.2 is used to find the model $\mathcal{M}_1$ maximum likelihood shape parameters α_{ML1} for that layer. This parameter set is then used to initialise the search for the set α_{ML} for each of the models $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ in turn. Model selection is then performed using 5 measures: Occam Factors (OF), maximum likelihood (ML), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and MAP likelihood evaluation. Results are given in Figure 7.

Model $\mathcal{M}_1$: Rectangle

Because the layer of the door is well represented by the rectangle model with 6 parameters, the maximum likelihood parameters for more complex models are the same, and the ML measure is not altered for different models. Each other measure selects $\mathcal{M}_1$ because it includes the fewest parameters.

Model $\mathcal{M}_2$: Arch

Models $\mathcal{M}_1$ and $\mathcal{M}_3$, which do not contain arches, are clearly inadequate for this layer. The maximum likelihood of models $\mathcal{M}_2$ and $\mathcal{M}_4$ is very similar, so again the maximum likelihood measure is ambiguous while the other measures all select the simpler model $\mathcal{M}_2$.

Model $\mathcal{M}_3$: Bevelled Rectangle
Models $\mathcal{M}_1$ and $\mathcal{M}_2$, which do not incorporate bevelling, fit the indentation poorly at its sloped edges. Models $\mathcal{M}_3$ and $\mathcal{M}_4$ have similar likelihoods, so $\mathcal{M}_3$ is chosen by all measures except ML.
Model $\mathcal{M}_4$: Bevelled Arch
In this case only the most complex model adequately describes the data. It is chosen by all measures depsite its complexity, as the likelihood of the data for this model is significantly higher than for other models.

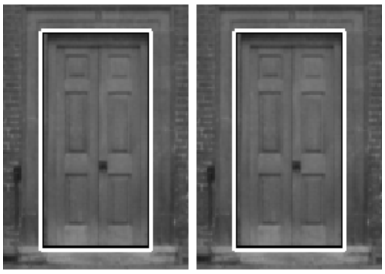
For each model, each model selection measure is clearly dominated the likelihood term. The OF, BIC, AIC and MAP measures all give similar results and appear adequate for preventing model overfitting. However the Occam factor is more theoretically sound than the other measures and incurs little extra computational expense, and is therefore preferred.

5.2 Model selection for the hyperparameters

Having selected a shape model for each layer in the scene, it is possible to discern not only between individual shapes, but also between configurations of shapes. As a simple test case, the evidence for a set of layers having no geometric alignment ($p(\beta)$ uniform, and $p(\alpha|\beta)$ uniform) is compared with the evidence for their belonging to a row of primitives, with priors given in Table 2. In this section, evidence is measured only using the Occam factor.
Gateway scene:
Figure 8(a) gives the layer models selected for each layer in the scene. In Figure 8(b) the evidence for each combination of two or more layers belonging to a row of identical primitives (black bars) is compared to the evidence for their being a uniformly distributed collection of shapes (white bars). No prior preference is expressed for either of these models. Clearly any combination of layers which includes the gateway is more likely to be part of a random scene, as the gateway is quite dissimilar in size and shape to the indentations. However the evidence for the two indentations taken by themselves belonging to a row is much higher than for their belonging to a general structure. Having detected this regularity, the indentations can be represented using 8 parameters (7 for one indentation, and the x position of the other) rather than 14. If such regularity can be detected in several collections of shapes, it can in turn be used to form hypotheses about higher level structure, such as the architectural style of the building as a whole.
Gothic church scene:
Figure 8(d) gives the evidence for several combinations of layers from the segmentation in Figure 8(c) belonging to a row as opposed to an arbitrary structure. There is a clear preference for a model with no regularity when the layers chosen include both windows and columns. However when only the windows are tested, the row model is clearly preferred, depsite the window parameters being slightly different due to some fitting errors caused by ambiguity near the boundary of each layer. Similarly the row model was preferred for the three columns, again allowing both a compact representation of the scene and the possibility of higher level inference about the scene strucure.

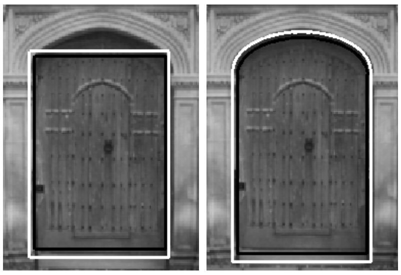
Model $\mathcal{M}_1$: Rectangle

Measure	Model			
($\times 10^4$)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
OF	**1.4781**	1.4787	1.4785	1.4797
ML	**1.4731**	**1.4731**	**1.4731**	**1.4731**
BIC	**1.4764**	1.4775	1.4775	1.4786
AIC	**1.4743**	1.4745	1.4745	1.4747
MAP	**1.4750**	1.4753	1.4753	1.4757



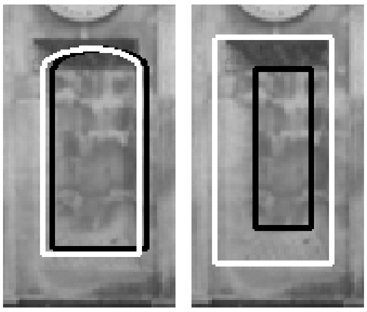
Model $\mathcal{M}_2$: Arch

Measure	Model			
($\times 10^4$)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
OF	2.5870	**2.5678**	2.5872	2.5691
ML	2.5817	**2.5618**	2.5807	**2.5618**
BIC	2.5850	**2.5657**	2.5842	2.5662
AIC	2.5829	**2.5632**	2.5821	2.5634
MAP	2.5837	**2.5641**	2.5829	2.5643



Model $\mathcal{M}_3$: Bevelled Rectangle

Measure	Model			
($\times 10^3$)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
OF	8.2704	8.2789	**8.1750**	8.1785
ML	8.2201	8.2169	**8.1139**	**8.1139**
BIC	8.2536	8.2453	**8.1524**	8.1582
AIC	8.2213	8.2183	**8.1153**	8.1155
MAP	8.2385	8.2378	**8.1353**	8.1376



Model $\mathcal{M}_4$: Bevelled Arch

Measure	Model			
($\times 10^4$)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
OF	2.3682	2.3623	2.3588	**2.3512**
ML	2.3628	2.3561	2.3528	**2.3444**
BIC	2.3661	2.3599	2.3566	**2.3488**
AIC	2.3640	2.3575	2.3542	**2.3460**
MAP	2.3649	2.3586	2.3552	**2.3472**

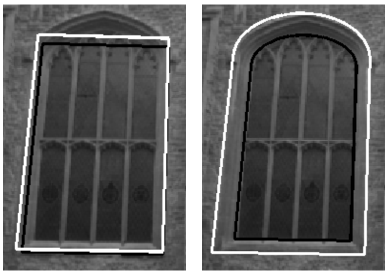


Fig. 7. *Evidence evaluation for single shapes. From left to right in each row: negative log evidence for this shape being an instance of each shape model, worst fit shape, best fit shape. Occam factor (OF), Maximum likelihood (ML), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and MAP probability measures are given. The model selected by each measure is in bold face.*

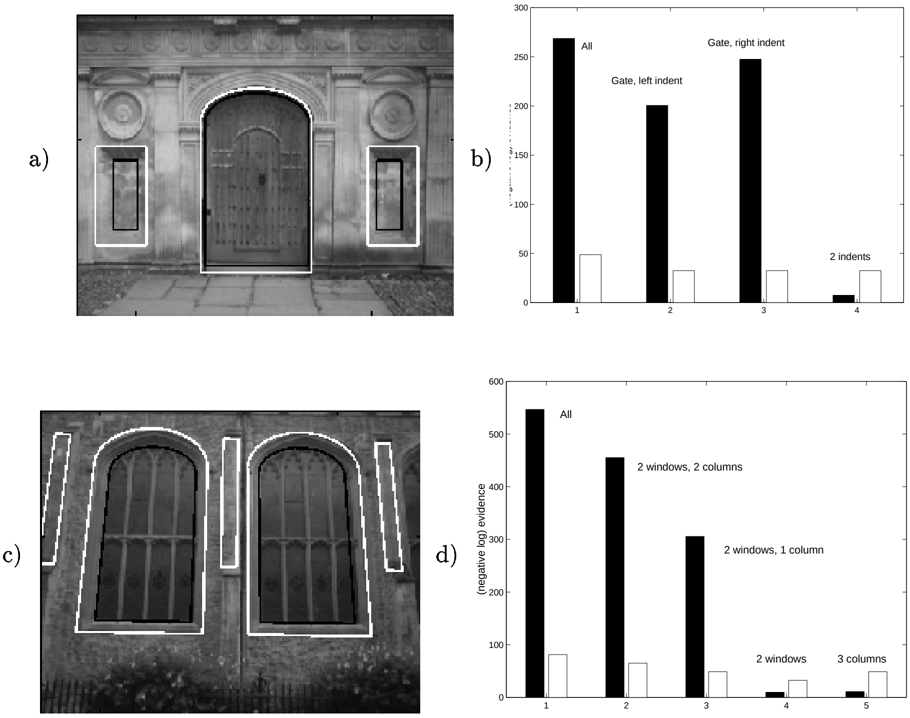


Fig. 8. *Testing for rows of similar shapes. The black bar is the (negative log) evidence for the shapes belonging to the row model; the white bar is the evidence for shapes having no regularity (see Section 5.2). Evidence has been normalised by subtracting out common factors.*

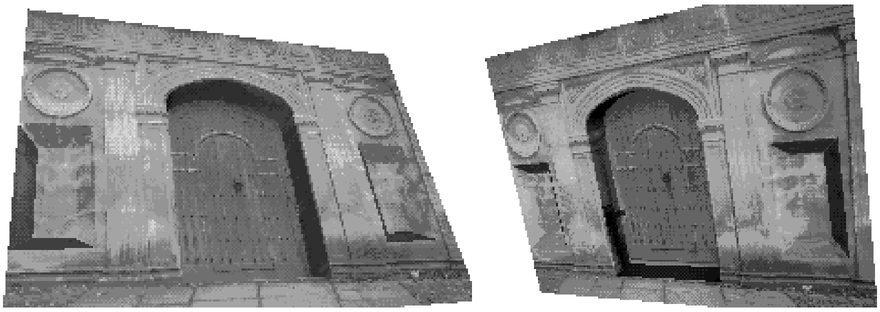


Fig. 9. *Recovered 3D surface of the Caius gateway scene.*

5.3 Selecting the number of layers

Comparison of the evidence can determine the number of layers present in a scene as well as their shape. Figure 10 gives the evidence for the gateway scene being modelled by 3, 2 and 1 primitives, which is clearly maximised for the 3 primitive case. The subsequent addition of a spurious primitive such as a depth 0 rectangle decreases the Occam factor while the likelihood remains constant, and hence is not selected.

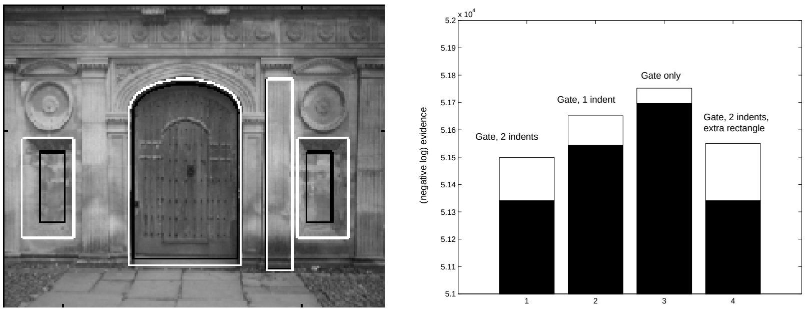


Fig. 10. *Negative log evidence for different numbers of layers in the gateway scene. From left to right: evidence for 3 detected layers, evidence for the gateway and only one indentation, evidence for the gateway only, evidence for all layers including spurious rectangle (shown above). The 3 and 4 layer models are clearly preferred to those with 1 and 2 layers; the 3 layer model is selected as it has a higher Occam factor.*

6 Conclusion

This paper presents a novel approach to layer extraction with the aim of creating a 3D model of the images that accurately reflects prior belief. This has been effected by a Bayesian approach with explicit, rather than implicit modelling of the distribution over segmentations. Given a hypothesised segmentation it is shown how to evaluate its likelihood and how to compare it with other hypotheses. A variety of model selection measures are considered, all but the most basic of which prove adequate to prevent model overfitting for the architectural scenes on which this approach is demonstrated. The Occam factor is recommended as is more accurate and theoretically sound while incurring minimal extra computational cost.

The hierarchical nature of the shape model means that it is easily extended to more complex scenes than those presented here. Future work will extend the number and type of shape primitives modelled, and the number of levels in the hierarchical shape model. For example, it should be possible to infer both the minimal parametrisation and the architectural style (e.g. Gothic, Georgian, modern bungalow) of the scene. A fully automatic initialisation scheme will use these hierarchical models to constrain an initial search for primitives in each image.

References

1. S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *International Conference on Computer Vision*, pages 777–784, 1995.
2. S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *IEEE Computer Vision and Pattern Recognition*, pages 434–441, 1998.
3. R. Cipolla, D. Robertson, and E. Boyer. Photobuilder – 3d models of architectural scenes from uncalibrated images. In *IEEE Int. Conf. on Multimedia Computing and Systems*, 1999.
4. A.R. Dick and R. Cipolla. Model refinement from planar parallax. In *Proc. 10th British Machine Vision Conference (BMVC'99)*, volume 1, pages 73–82, Nottingham, 1999.
5. A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis.* Chapman and Hall, Boston, 1995.
6. U. Grenander, Y. Chow, and D.M. Keenan. *HANDS. A Pattern Theoretical Study of Biological Shapes.* Springer-Verlag. New York, 1991.
7. F. Heitz and P. Bouthemy. Multimodal motion estimation and segmentation using markov random fields. In *Proc. 10th Int. Conf. Pattern Recognition*, pages 378–383, 1991.
8. A. Jepson and M. Black. Mixture models for optical flow computation. In *IEEE Computer Vision and Pattern Recognition*, pages 760–766. IEEE, 1993.
9. R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In *European Conference on Computer Vision*, pages 55–71, 1998.
10. D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.
11. J. Magarey and N. Kingsbury. Motion estimation using a complex-valued wavelet transform. *IEEE Trans. Signal Processing*, 46(4):1069–1084, April 1998.
12. M. Pollefeys, R. Koch, and L. VanGool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *International Conference on Computer Vision*, pages 90–95, 1998.
13. G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
14. S. Seitz and C. Dyer. A theory of shape by space carving. In *International Conference on Computer Vision*, pages 307–314, 1999.
15. H.Y. Shum, M. Han, and R. Szeliski. Interactive construction of 3d models from panoramic mosaics. In *IEEE Computer Vision and Pattern Recognition*, pages 427–433, 1998.
16. D. S. Sivia. *Data Analysis: A Bayesian Tutorial.* Oxford University Press, Oxford, 1996.
17. P. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. In *International Conference on Computer Vision*, pages 983–990, 1999.
18. J. Wang and E. H. Adelson. Layered representation for motion analysis. In *IEEE Computer Vision and Pattern Recognition*, pages 361–366, 1993.
19. T. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation. In *IEEE Computer Vision and Pattern Recognition*, pages 321–326, 1996.