An Integrated Bayesian Approach to Layer Extraction from Image Sequences

Philip H.S. Torr, Richard Szeliski, and P. Anandan

Abstract—This paper describes a Bayesian approach for modeling 3D scenes as a collection of approximately planar layers that are arbitrarily positioned and oriented in the scene. In contrast to much of the previous work on layer-based motion modeling, which computes layered descriptions of 2D image motion, our work leads to a 3D description of the scene. There are two contributions within the paper. The first is to formulate the prior assumptions about the layers and scene within a Bayesian decision making framework which is used to automatically determine the number of layers and the assignment of individual pixels to layers. The second is algorithmic. In order to achieve the optimization, a Bayesian version of RANSAC is developed with which to initialize the segmentation. Then, a generalized expectation maximization method is used to find the MAP solution.

Index Terms—Layer extraction, segmentation, stereo matching, motion estimation.

1 Introduction

Extracting three-dimensional models from a series of still images (the structure from motion problem) has been one of the defining problems for computer vision involving geometry, segmentation, and probability theory. Three-dimensional scene modeling from multiple images can be broken into two subproblems: creating a 3D geometric model of the scene and creating a texture map that captures the visual appearance of the scene. Despite a large amount of research into the area (a summary of which can be found in [8]), the recovery of accurate depth information remains only partially solved. Most existing algorithms perform reasonably well when when matching detected features or in textured regions, but perform poorly around occlusion boundaries and in untextured regions.

One approach that is somewhat effective in mitigating these two problems is the layered approach [1], [17], [18]. Here, the aim is to segment the image into a set of regions, such that pixels within each region move in a manner consistent with a 2D parametric transformation (e.g., affine). Then, the edges of the region correspond to occlusion boundaries and the motion of all the pixels within the region (textureless or not) is determined by the (six in the case of affine) parameters of the motion model. Typically, the EM (expectation maximization) algorithm [7] is used to effect a segmentation. The use of the 2D affine motion model corresponds to an assumption that the layer maps to a plane in 3D viewed under orthographic conditions. This assumption is valid for a wide variety of scenes, even for nonplanar objects for which the distance to the camera is sufficiently large relative to the depth variation across the object. The layered method tends to fail when nonplanar objects are viewed in close up as the representation is not adequate in this case. In fact, any object that contains significant parallax effects will cause the breakdown of a layered representation with a global motion model. In order to overcome this problem, we propose an approach in which each pixel within a layer can have an associated disparity. The 2D layers, are usually not intended to capture 3D scene structure. In contrast, it was proposed in [2] that the scene should be decomposed into a collection of *3D layers* (or *sprites*), each of which consists of a plane equation, a color image that captures the appearance of the sprite, a per-pixel opacity map, and a per-pixel depth-offset relative to the nominal plane of the layer. The advantage of this approach is that roughness or parallax effects on the layer can be modeled without the overfragmentation and instability inherent in a purely 2D parametric approach. This approach to layered representation can be viewed as an extension of the plane plus parallax decomposition of image disparities across multiple views. The class of scenes the method works well for includes those that the 2D layered method works well for, but pushes the envelope to include scenes for which there is a small amount of per pixel parallax on each plane. Although Baker et al. [2] proposed a generative model for images based on layers, the algorithm described for recovering the layers from a given set of input images is incomplete. Layer initialization was based *solely* on manual user input and, as the paper confesses, the final assignment of pixels to layers was not fully developed.

In fact, the central problems of layer-based scene modeling are the determination of the number of layers and the assignment of pixels to layers. An infinite number of decompositions are equally consistent with the generative model proposed in [2], ranging from assigning one layer plane to every pixel (with no depth offset) to modeling the entire scene with a single layer plane (with lots of depth offset). Apart from object and scene level semantic information (which is important, but outside the scope of this paper), the natural criterion to use is compactness or *parsimony* of description. But computing a parsimonious description is implicitly and tightly tied to our prior assumptions about the scene and the imaging process that generated the input images. Bayesian decision theory provides a foundation for formally stating our prior assumptions and then developing algorithms for applying these priors during reconstruction.

2 The Bayesian Method

The approach that follows is unashamedly Bayesian, to quote E.T. Jaynes: “Vision is *Bayesian* inference from incomplete data” [12], in contrast to the mechanical approach set out in [6] or a purely geometric approach [13]. As set out in his Internet book [12] (which every vision postgraduate would be well advised to read), the Bayesian method provides a consistent way of reasoning about the world that can be viewed as an extension of the Aristotelian calculus of logic to uncertainty. The Bayesian approach was probably first proposed for vision (at least for segmentation) by Besag [5] and then elaborated by Geman and Geman [10]. The use of least committed priors for the world was set out in Grenander et al. [11] and Bayesian stereo was developed by Szeliski [15] and Belhemeur and Mumford [3] among others.

Some will complain that to use Bayesian methods one must introduce arbitrary priors on the parameters. However, far from being a disadvantage, this a tremendous advantage as it forces open acknowledgment of what assumptions were used in designing the algorithm, which all too often are hidden away beneath the veneer of equations. Furthermore, there is nothing wrong with injecting our prior information into the design of vision algorithms—all vision researchers do it whenever they choose the probability distribution of their error. As will be seen, the validity of the prior assumptions can be tested by altering the priors and seeing whether this leads to radically different results.

In addition to the utilization of prior information, Bayesian methods are further distinguished from orthodox statistical

- *P.H.S. Torr is with Microsoft Research Ltd., Microsoft Research, St. George House, 1 Guildhall St., Cambridge CB2 3NH, UK. E-mail: philtorr@microsoft.com.*
- *R. Szeliski and P. Anandan are with Microsoft Research, One Microsoft Way, Redmond WA 98052. E-mail: {szeliski, anandan}@microsoft.com.*

Manuscript received 22 Feb. 2000; revised 24 Oct. 2000; accepted 12 Nov. 2000.
Recommended for acceptance by H. Christensen.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 111542.

0162-8828/01/$10.00 © 2001 IEEE

methods by their use of full probability distributions rather than the mode to describe parameters. This difference can be illustrated by comparing two competing models for the data. This will be useful later when we need to determine the number of layers used to describe the image.

Given a closed set[1] of k models $\mathbf{M}_1 \ldots \mathbf{M}_k \in \mathcal{M}$ that can explain the data $\mathbf{D}$ Bayes rule leads us to

$$\Pr(\mathbf{M}_i|\mathbf{DI}) = \frac{\Pr(\mathbf{D}|\mathbf{M}_i\mathbf{I})\Pr(\mathbf{M}_i|\mathbf{I})}{\Pr(\mathbf{D}|\mathbf{I})},$$

where $\mathbf{I}$ is the prior information assumed about the world. In this paper, the set of models are: $\mathbf{M}_1$ that the data can be explained by one layer, $\mathbf{M}_2$ that the data can be explained by two layers, etc. The posterior probability that a given model $\mathbf{M}_i$ is the correct one is

$$\Pr(\mathbf{M}_i|\mathbf{DI}) = \frac{\Pr(\mathbf{D}|\mathbf{M}_i\mathbf{I})\Pr(\mathbf{M}_i|\mathbf{I})}{\sum_{j=1}^{j=k}\Pr(\mathbf{D}|\mathbf{M}_j\mathbf{I})\Pr(\mathbf{M}_j|\mathbf{I})}. \quad (1)$$

Note that by construction, $\sum_{j=1}^{j=k}\Pr(\mathbf{M}_j|\mathbf{D}) = 1$. The key to this equation is the evaluation of $\Pr(\mathbf{D}|\mathbf{M}_j\mathbf{I})$ which is called the evidence. In contrast to non-Bayesian techniques, the evidence for a model is in fact the integral of the likelihood over all possible values of the model's parameters:

$$\Pr(\mathbf{D}|\mathbf{M}_j\mathbf{I}) = \int \Pr(\mathbf{D}|\mathbf{M}_j\boldsymbol{\theta}_j\mathbf{I})\Pr(\boldsymbol{\theta}_j|\mathbf{M}_j\mathbf{I})\partial\boldsymbol{\theta}_j \quad (2)$$

$$\text{Evidence} = \int \text{likelihood} \times \text{prior}\ \partial\boldsymbol{\theta}_j, \quad (3)$$

where $\boldsymbol{\theta}_j$ are the kth model's parameters, and $\Pr(\boldsymbol{\theta}_j|\mathbf{M}_j\mathbf{I})$ is the prior distribution of parameters of the model. How does this relate to the principle of parsimony so prevalent in model selection (AIC, MDL, etc.)? Not as it commonly believed, i.e., by penalizing more complex models a priori.[2] In the absence of any prior preference between the two models, $(\Pr(\mathbf{M}_i) = \Pr(\mathbf{M}_j))$, all the models are equally likely a priori. Rather, more complex models have the probability of the prior dispersed over a greater region of parameter space. Note by definition $\int \Pr(\boldsymbol{\theta}_j|\mathbf{M}_j\mathbf{I})d\boldsymbol{\theta}_j = 1$. Thus, a more complex model will only be supported if there is a corresponding increase in the likelihood of the data given that model. This is a subtle point and worthy of some reflection. Equation (3), representing the evidence for a given model, is a crucial equation and is worthy of close scrutiny. **If one accepts Bayes rule then logic dictates that this is the only way to calculate the posterior probability of each model [12].**

How to use this information in selecting the most appropriate model requires a little more machinery which is furnished by Bayesian Decision Theory. We have a set of potential actions $a_1 \ldots a_k \in \mathcal{A}$, each corresponding to selecting one of the models $\mathbf{M}_1 \ldots \mathbf{M}_k$. To decide on a course of action, a cost/loss function $u() : \mathcal{A} \times \mathcal{M} \rightarrow \Re$ must be defined attaching utilities to each action $a \in \mathcal{A}$ given a state of the world $M \in \mathcal{M}$. The optimal action is that which maximizes the expected utility: $\max_a \sum_i u(a, \mathbf{M}_i)\Pr(\mathbf{M}_i|\mathbf{DI})$. The simplest case is the zero-one cost function that provides a reward of one in the case of the correct model and zero if the choice is incorrect. This choice of cost leads to selecting the model that maximizes the posterior. This is the loss function that we consider in this paper, noting that others may be appropriate depending on the application.

Another useful Bayesian technique is that of marginalization, which allows us to dispose of parameters that are not directly useful to us at any given time by integrating them out. This is effected by the following identity: $\int_{-\infty}^{\infty}\Pr(\mathbf{X},\mathbf{Y}|\mathbf{I})\partial\mathbf{Y} = \Pr(\mathbf{X}|\mathbf{I})$. As it turns out, the Bayesian technique of marginalization allows us to finesse a problem that has plagued motion algorithms for decades, namely the correspondence problem.

3 First Stage—Preprocessing

The input to the algorithm is a sequence of images. These are the (raw) data $\mathbf{D}$. The first stage of the algorithm uses a coarse to fine algorithm to obtain initial disparity estimates for each pixel [4]. Then, calibration and camera matrices are recovered [16]. Each image is then transformed in such a way that the plane at infinity is stabilized. In other words, the images are rectified to the first one by transforming under $\mathbf{H}_\infty^{1j}$ the homography of the plane at infinity between image 1 and the jth image to remove the effects of rotation. Registering each image to the plane at infinity has the effect that the disparity (motion) of each pixel now depends purely on depth. Here, the disparity $\delta(x, y)$ at pixel (x, y) in image 1 is taken to mean the motion along the epipolar line between image 1 and image 2. Because the images are rectified to the plane at infinity, it is a bijective function of the depth of that pixel, $\delta(x, y) = \rho(Z(x, y))$. The transformation ρ is purely a function of the calibration and camera matrices. The notation $\mathbf{Z}$ is adopted for the set of depths and $Z(x, y)$ for the depth of a pixel (x, y) in the first image.

4 Parametric Formulation of Problem

The set of input images is denoted by $\mathbf{D}$ (the data), the model $\mathbf{M}$, consisting of a set of m planes $\boldsymbol{\Theta}$ with parameters $\boldsymbol{\theta}_j$, $j = 1 \ldots m$, and a set $\mathbf{L}$ of per pixel labels $l(x, y)$ for the first image. Nothing is known about the number of layers m—this must also be recovered. One image (image 1) is used to initialize the segmentation and from here on the segmentation and labeling is done in the coordinate system of image 1. Because interimage motions are small, it is more natural to do the segmentation in the image rather than in some 3D-based coordinate system. Hence, the aim is simply to extract $\mathbf{M}$ from $\mathbf{D}$.

The parameters of each plane are $\boldsymbol{\theta} = (a, b, c)$ such that $aX + bY + cZ = 1$, where X, Y, Z are the Euclidean coordinates. This parametrization is chosen as it excludes all planes passing through the origin of the coordinate system (the optic center of the first camera) i.e., planes of the form $aX + bY + cZ = 0$. These project to a line in the first image (and subsequent images if the baseline is small) and, thus, correspondence cannot be recovered for them. Note that $\boldsymbol{\theta} = (a, b, c)$ lies along the normal of the plane. The coordinate system is chosen such that the origin is at the first camera's optic centre. In image 1, $x = X/Z$ and $y = Y/Z$, leading to $ax + by + c = 1/Z$. Thus, given the plane and the (x, y) coordinate of any pixel, its depth may be found (and, hence, its corresponding pixel in any other image). For the case when the direction of motion along the optic axis is small relative to the distance to the 3D point, $1/Z$ is roughly proportional to the disparity between images. The coefficients a and b give the disparity gradients in the x, y directions and c gives the inverse depth of the plane at the principal point.

One plane is privileged in that it is always represented by a layer and has fixed parameters $\boldsymbol{\theta}_\infty = (a, b, c) = (0, 0, 0)$: This is the plane at infinity. Although this ideal cannot truly exist in a Euclidean representation, it serves a useful purpose. Pixels that are so distant that their disparity is swamped by noise (e.g., sky) have very ill-conditioned depths and cannot be easily segmented. These are all grouped together into the plane at infinity.

[1] This is an important prerequisite and the burden is on us to explore the most likely models.

[2] Although there is nothing to stop us doing this, it is not necessary.

5 Probabilistic Formulation

This section introduces the mixture model used to describe the layers. Although some will consider this standard fare, we consider it worth looking under the bonnet[3] to show the logical implications of our formulation. We will show how Bayesian reasoning can be used to describe two vision heuristics of long standing: the *disparity gradient limit* and *plane plus parallax*. We will also show how the layers can be recovered without estimating correspondence. It transpires that this is very useful for segmentation, as it finesses the problem of mismatches.

5.1 Posterior Probability of the Model

The model parameters $m, \mathbf{\Theta}, \mathbf{L}$ are chosen so as to maximize the posterior probability:

$$\max_{m\mathbf{\Theta}\mathbf{L}} \Pr(m\mathbf{\Theta}\mathbf{L}|\mathbf{D}\mathbf{I}) = \frac{\Pr(\mathbf{D}|m\mathbf{\Theta}\mathbf{L}\mathbf{I})\Pr(m\mathbf{\Theta}\mathbf{L}|\mathbf{I})}{\Pr(\mathbf{D}|\mathbf{I})}. \tag{4}$$

The denominator can be discounted for the purposes of parameter estimation, (but not for the purposes of model comparison as will be seen later), as it is constant for all values of the parameters. So far, the estimation of depth (or disparity) has not been mentioned, although it would apparently have a direct bearing on the likelihood.

5.2 Recovering the Plane Parameters without Correspondence

We can use the Bayesian method of marginalization to remove the depth parameter from the posterior probability of the plane

$$\Pr(m\mathbf{\Theta}\mathbf{L}|\mathbf{D}\mathbf{I}) = \int_{\mathbf{Z}} \Pr(m\mathbf{\Theta}\mathbf{L}\mathbf{Z}|\mathbf{D}\mathbf{I})\partial\mathbf{Z}. \tag{5}$$

Later, it will be seen that this is most convenient when trying to determine what label a pixel should have, or when reestimating the planes using generalized expectation-maximization [9] GEM described below. The advantage of marginalization is that it allows us to use a plane to capture the motion of a region of an image, but also allows for relief (or parallax) out of that plane. By marginalizing the depths rather than doing full MAP estimation, we avoid a strong commitment to depth estimation. Typically, too early a commitment to a depth estimate may result in convergence to a local (rather than global) maximum of the posterior-likelihood (especially in homogeneous regions of the image). By marginalizing, we are in effect "hedging our bets" and not committing to a single depth estimate. Rather, the distribution of depth at a pixel is specified by the mixture model. Next, the posterior-likelihood will be decomposed into its component parts and it will be explained how it can be optimized using the GEM algorithm.

5.3 Decomposition

Assuming that the number of layers m has been determined (techniques to do this are set out below) and the noise across the image is not spatially correlated, the posterior-likelihood can be evaluated as the product of the MAP likelihoods at each individual pixel.

$$\int_{\mathbf{Z}} \Pr(\mathbf{\Theta}\mathbf{Z}\mathbf{L}|\mathbf{D}\mathbf{I})\partial\mathbf{Z} \propto \int_{\mathbf{Z}} \prod_{xy} \Pr(\mathbf{D}|\mathbf{\Theta}z(x,y)l(x,y)\mathbf{I})\Pr(\mathbf{\Theta}\mathbf{L}\mathbf{Z}|\mathbf{I})\partial\mathbf{Z}. \tag{6}$$

Next, let us consider each pixel individually, drop the (x, y) index, adopt the notation l_j for $l(x, y) = j$, and let $\tilde{\mathbf{L}}$ be the set of labels excluding the label for pixel (x, y). Then,

$$(\mathbf{D}|\mathbf{\Theta}lz\mathbf{I}) = \begin{cases} (\mathbf{D}|\mathbf{\Theta}_i\mathbf{I}) & \text{if } l(x,y) = i \\ 0 & e \geq T, \end{cases}$$

which is a mixture model [9] between the layers, with spatial correlation between the label parameters.

5.4 The Likelihood

The term $\Pr(\mathbf{D}|z\boldsymbol{\theta}_j\mathbf{I})$ is the likelihood of the pixel having a depth (or disparity) hypothesis z. It can be evaluated from the cross-correlation between the pixel in question and its correspondences in each other image of the sequence. As such, it only depends directly on the depth, and fixing z can be written $\Pr(\mathbf{D}|\boldsymbol{\theta}_j z\mathbf{I}) = \Pr(\mathbf{D}|z\mathbf{I})$. (This is logically correct; the likelihood purely depends on the estimated disparity, it cannot depend on anything else. How z is influenced by $\boldsymbol{\theta}_j$ is explained below.) Suppose that the variation in intensity between images can be modeled as Gaussian with mean μ_i and standard deviation σ_i. Let $\Delta i_j(x, y)$ be the difference in (color) intensity between the pixel in image 1 and its corresponding pixel in image j. Then,

$$\Pr(\mathbf{D}|z\mathbf{I}) = \prod_{j \neq 1}\big((1 - p_o)\Phi(\Delta i_j(x,y)|\mu_i\sigma_i) + \alpha p_o\big),$$

where $\Phi(\Delta i_j(x, y)|\mu_i\sigma_i)$ is the Gaussian-likelihood

$$\Phi(\Delta i_j(x,y)|\mu_i\sigma_i) = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)\exp -\frac{\Delta i_j(x,y)}{2\sigma_i^2}$$

and p_o is the probability of occlusion, or that the pixel is in some other way radically different (for instance, due to the interpolation error when working out the cross-correlation), and α is a constant being the probability of the intensity difference given an occlusion (uniform over the range of intensity). Equation (5) is a form of contaminated Gaussian with parameters $\mu_i, \sigma_i, \alpha, p_o$ and distribution denoted by $\Upsilon(\Delta i_j|\mu_i, \sigma_i, \alpha, p_o)$. It provides a robust error metric: The effect of any one observation is bounded. In our work, $p_o = 0.05$ (although there is a switch in the code to switch this up between $0.05 - 0.1$ depending on how far away the pixel is, encoding the fact that more distant pixels tend to be occluded more), and $\alpha = (1/256)^3$, the range of pixel intensities for RGB.

As mentioned earlier, the depth is integrated out. To do this, the likelihood is discretized. To discretize the likelihood given, for each pixel, the likelihood (5) is estimated over a set of disparity hypotheses. Typically, the scenes that we are dealing with are from video sequences, the interframe motion is 0-4 pixels, thus 20 disparity hypotheses increasing in steps of 0.2 can be used to sample the 0-4 pixel disparity range. Next, the form of the priors are explained.

5.5 The Priors

Using the product rule, the prior can be decomposed as follows: $\Pr(\mathbf{Z}\mathbf{\Theta}\mathbf{L}|\mathbf{I}) = \Pr(\mathbf{Z}|\mathbf{\Theta}\mathbf{L}\mathbf{I})\Pr(\mathbf{\Theta}\mathbf{L}|\mathbf{I})$. There is no reason to assume a prior correlation between the parameters and shape of the projection of a plane[4] thus, $\Pr(\mathbf{\Theta}\mathbf{L}|\mathbf{I}) = \Pr(\mathbf{L}|\mathbf{I})\Pr(\mathbf{\Theta}|\mathbf{I})$. The prior $\Pr(\boldsymbol{\theta}|\mathbf{I})$ on a given plane's parameters is assumed to be Gaussian on the parameters a, b, c with zero mean and standard deviations σ_a, σ_b, and σ_c. This has a very interesting physical interpretation. Since a and b represent the disparity gradients, σ_a and σ_b can be chosen to favor fronto-parallel planes and to control the disparity gradient limit of the plane. This elegantly combines Bayesian reasoning with an old vision heuristic. The parameter σ_c is a weak prior favoring more distant planes and penalizes ones that are too close to the camera. These are weak priors and will be overruled by observed data. They serve as a regularization that helps finesse the effects of outliers and ambiguity.

[3] Hood.

[4] This is not entirely true as one would expect distant objects to be smaller in extent, but we do not consider that here.

The prior $\Pr(\mathbf{Z}|\mathbf{\Theta}\mathbf{L}\mathbf{I})$ controls the amount of parallax favored. In real situations, points will not always lie exactly on a plane. Yet, many surfaces can be modeled as a plane together with some relief leading to the much vaunted plane plus parallax algorithms. However, this idea is typically used as a heuristic without concrete definition. Bayesian methods allow the idea to be made concrete, by defining the distribution of $\Pr(\mathbf{Z}|\mathbf{\Theta}\mathbf{L}\mathbf{I})$ in terms of a distribution of the parallax from the plane. This allows the plane to be recovered without knowing the exact disparities. The distribution $\Pr(\mathbf{Z}|\mathbf{\Theta}\mathbf{L}\mathbf{I})$ is specified in terms of the amount of parallax, as a mean zero Gaussian with $\sigma_p = 0.5$. This may then be convolved with the discretized-likelihood specified above. To recover the likelihood that any given pixel belongs to a given layer j, given the plane parameters $\boldsymbol{\theta}_j$, the integrated-likelihood may be used:

$$\Pr\big(\mathbf{D}|l_j\mathbf{I}\big) = \int_z \Pr(\mathbf{D}|z\mathbf{I}) \Pr\big(l_j|\mathbf{I}\big) \Pr\big(z|\boldsymbol{\theta}_j\mathbf{I}\big) \Pr\big(\boldsymbol{\theta}_j|\mathbf{I}\big) \Pr\big(l_j|\tilde{\mathbf{L}}\mathbf{I}\big)\partial z. \tag{8}$$

A uniform prior distribution is taken on z. This is easier than it looks to evaluate, since the integration merely involves 20 multiplications, one for each putative disparity.

The prior $\Pr(\mathbf{L}|\mathbf{I})$ represents our belief about the likelihood of the spatial disposition of the world. In the general case, it is not known how to evaluate this. Here, we use an MRF (Markov Random Field) formulation [5], [10]. Therefore, what can be evaluated is the probability that pixel (x, y) has a label k given a local neighborhood in $\mathbf{L}$. What follows is the update rules for a single label given its neighbors. Let $l_k(x, y)$ be an indicator variable such that $l_k(x, y) = 1$ if pixel (x, y) is in the kth layer, or 0 otherwise. Then,

$$\Pr\big(l_k(x,y)|\tilde{\mathbf{L}}\mathbf{I}\big) = \frac{\Pr\big(\tilde{\mathbf{L}}|l_k(x,y)\mathbf{I}\big)\Pr\big(l_k(x,y)|\mathbf{I}\big)}{\Pr\big(\tilde{\mathbf{L}}|\mathbf{I}\big)}.$$

The normalizing constant is just

$$\Pr\big(\tilde{\mathbf{L}}|\mathbf{I}\big) = \sum_{j=1}^{j=m} \Pr\big(\tilde{\mathbf{L}}|l_j(x,y)\mathbf{I}\big)\Pr\big(l_j(x,y)|\mathbf{I}\big). \tag{9}$$

The prior $\Pr(l_k(x, y)|\mathbf{I})$ is simply the probability that a given pixel lies in a given layer. In the absence of other information, it seems reasonable that this should be uniform except, however, for the layer of the plane at infinity l_∞, which is deemed more likely a priori. Given points with low disparity (and, hence, high variance in Z) it is reasonable to assign them to the plane at infinity rather than some arbitrary and ill-conditioned plane. Next, we use a factored approximation $\Pr\big(\tilde{\mathbf{L}}|l_k(x,y)\mathbf{I}\big) \approx \prod_{uv} \Pr(l(u,v)|l_k(x,y)\mathbf{I})$. As $l(u, v)$ is not known (only its distribution is known), the above quantity is replaced by its expectation when using EM:

$$\Pr\big(\tilde{\mathbf{L}}|l_k(x,y)\mathbf{I}\big) \approx \prod_{uv}\sum_{j=1}^{j=m} \Pr\big(l_j(u,v)|l_k(x,y)\mathbf{I}\big)\Pr\big(l_j(u,v)|\mathbf{I}\big). \tag{10}$$

The question then is how to evaluate $p_{jk} = \Pr\big(l_j(u,v)|l_k(x,y)\mathbf{I}\big)$. What information do we have that might affect this distribution? All that we have a priori is the distance between the points Δd and the difference in their color values Δc. We would like the following properties for this distribution. If $l(u, v) = k$, we would like p_{jk} to be high. If the two pixels are close and/or of the similar color, they are more likely to have the same label, falling off to a probability $1/m$ (where m is the number of layers) if the pixels are far apart or dissimilar in color. We would like the converse to also be true: If $l(u, v) \neq k$, we would like p_{jk} to be low if the pixels have the same color or are near, rising to $m - 1/m$ if they are distant.

There is no clear answer for what this distribution should be. In the future, we hope to try and learn it from the data. In Section 8, we shall try several forms for the distribution. Here is one suggestion: The probability that the two pixels belong to the same layer $p_{jk}, j = k$ could be modeled by a contaminated Gaussian (defined above) $\Upsilon(\Delta c|\mu_c, \sigma_c, \alpha_c, p_c)$, where $p_c = 1/m$. The mixing parameter α_c controls the amount of homogeneity expected in the layer; the mean $\mu_c = 0$ and the standard deviation are set to be a function of the distance $\sigma_c = \beta_c/\Delta x$. This function satisfies all the desiderata given above, as well as possessing some interesting properties.

Consider the log probability that a given pixel has label k

$$\log\Pr\big(l_k(x,y)|\tilde{\mathbf{L}}\mathbf{I}\big) = \sum_{uv}\log(m\Pr(l(u,v)|l_k(x,y)\mathbf{I})) + \log\Pr(l_k(x,y)|\mathbf{I}) + \text{constant}. \tag{11}$$

For each pixel nearby that is expected to have label k, there will be a positive addition to this log-likelihood proportionate to the color similarity and inverse distance of that pixel. *In addition*, if neighboring pixels have a similar color but are likely to have a label other than k, there is a negative contribution to the log-likelihood. Thus, if a pixel takes on a particular interpretation, it not only excites its neighbors to have a similar interpretation, it also inhibits its neighbors of a similar color from having a different interpretation.

6 GENERALIZED EM

With the priors specified, the next problem is how to optimize the posterior-likelihood of the interpretation. One method of estimation that has been used successfully for estimation of mixtures is the EM algorithm [7] in which the labels are treated as missing data. EM is a useful procedure in finding the mode of a posterior distribution $\Pr(\Theta|\mathbf{D})$ in which it is hard to maximize $\Pr(\Theta|\mathbf{D})$ directly but easy to work with $\Pr(\Theta|\mathbf{L}\mathbf{D})$ and $\Pr(\mathbf{L}|\Theta\mathbf{D})$.

The EM algorithm proceeds as follows:

1. Estimate the number of layers m and the parameters of their associated planes using the algorithm described in Section 7.
2. Replace missing data values $\mathbf{L}$ by their expectations given the parameters $\boldsymbol{\theta}$.
3. Estimate parameters $\boldsymbol{\theta}$, assuming the missing data are given by their expected values.
4. Reestimate the missing values, assuming the new parameters are correct.
5. Reestimate the parameters, etc., iterating until convergence.

The EM algorithm has the very desirable property that each of its cycles will increase the posterior-likelihood.

6.1 E-Step

The expectation step proceeds as follows: For a given label $l_k(x, y)$, dropping terms that are independent,

$$\Pr\Big(\hat{l}_k|\mathbf{D}\boldsymbol{\theta}_k\tilde{\mathbf{L}}\mathbf{I}\Big) = \frac{\Pr(\mathbf{D}|l_k\mathbf{I})\Pr\big(\tilde{\mathbf{L}}|l_k\mathbf{I}\big)\Pr(l_k|\mathbf{I})}{\sum_{j=1}^{j=k}\Pr\big(\mathbf{D}|l_j\mathbf{I}\big)\Pr\big(\tilde{\mathbf{L}}|l_j\mathbf{I}\big)\Pr\big(l_j|\mathbf{I}\big)},$$

where the quantities on the right-hand side are those estimated at the previous iteration and $\hat{l}_k$ is to be estimated. This can be evaluated using (8) and (9).

6.2 M-Step

The maximization step involves finding the set of plane parameters Θ that maximize (4). This is computationally difficult if all of the plane parameters are to be maximized simultaneously as in

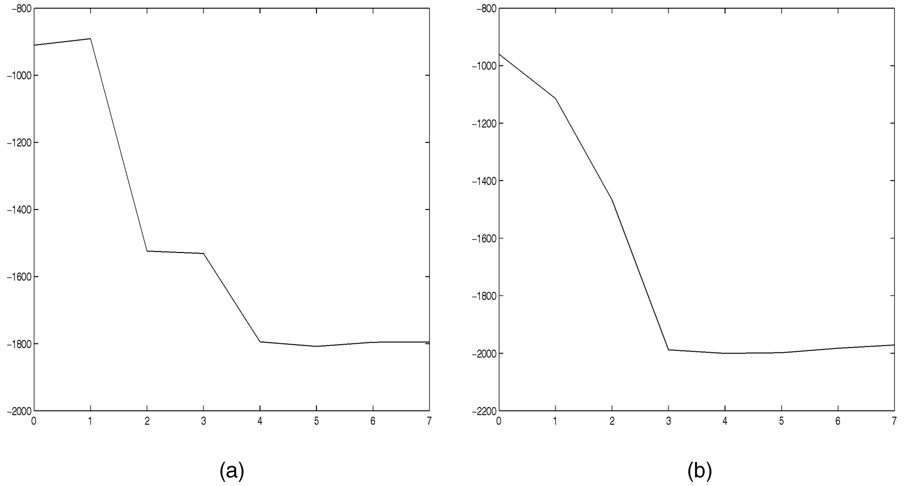

Fig. 1. The unnormalized approximation to the (negative log) posterior-likelihood of the number of layers for (left) the garden and (right) the Dayton sequences. For the garden sequence, it can be seen that after four layers are selected, the graph begins to increase with subsequent layers being less likely, but with no dramatic increase in posterior, even for seven layers. This is also the case for three layers in the Dayton example.

traditional EM, as there are $3m$ parameters to be determined, where m is the number of layers. Fortunately, we can use a generalized EM algorithm [9], in which the posterior-likelihood still increases at each iteration. Rather than maximize all the parameters simultaneously, subsets of the parameters are maximized in turn using a gradient descent technique, while the others are held constant. In this case, natural subsets are the parameters of each plane. The covariance matrix of each plane is approximated by the Hessian of the error function at the minimum.

7 Initialization

> "One forms provisional theories and waits for time or fuller knowledge to explode them. A bad habit, Mr. Ferguson, but human nature is weak."[5]

The GEM algorithm described in the last section is provable convergent to a maximum of the posterior distribution, which is all well and good, except that typically it will 1) take a very long time to converge and 2) converge to a local maximum. **The most important thing in a segmentation algorithm is the initialization**. The GEM algorithm is no panacea for poor initialization and for something as complicated as image segmentation, it will get trapped in local minima unless started with a good solution. Next, the PILLAGE algorithm is outlined for selecting the initial parameters of the planes.

7.1 Choosing the Number of Layers to Maximize the Posterior-Likelihood

The Bayesian method for model selection is encapsulated by (1) and the evaluation of the evidence for each model (one layer, two layers, three layers, etc.). Evaluation of the evidence involves integrating the posterior probability (4) over the prior range of the parameters. At present, this is simply not computationally feasible, so we must think of a good approximation to the evidence. Examination of the posterior probability (4) reveals that only pixels that have a high entropy distribution for $\Pr(\mathbf{D}|z\mathbf{I})$ (entropy is simply $\sum p \log p$, with normalization $\sum p = 1$) will affect the posterior distribution of Θ. This is simply the intuition that pixels for which there is great uncertainty about a correspondence (e.g., those within homogeneous regions) contribute little to the accuracy of the estimation of the plane parameters. Thus, computational effort is concentrated on those pixels with high entropy.

To detect these, a feature extractor is run on the first image and features with high entropy used as input to the algorithm. Next, for each model $\mathbf{M}_i$, with i layers, $i = 1 \dots k$, the $3i$ parameters of the plane Θ_i are robustly estimated from the high entropy points. To do the robust estimation, a RANSAC like algorithm is used (which shall be referred to as PILLAGE[6]), but with a vast improvement. Rather, select the sample that maximizes the number of inliers. The posterior (4) itself is estimated and the sample that maximizes this is selected as the initial estimate for GEM. In effect, PILLAGE is the Bayesian version of RANSAC.

The PILLAGE algorithm proceeds as follows:

1. Simultaneously sample 3 spatial close points for each plane.
2. Estimate the parameters of each plane $\boldsymbol{\theta}_i$, $i = 1 \dots k$.
3. Estimate label probabilities (one step of EM).
4. Calculate the posterior for this set of plane parameters.

The sampling is repeated a fixed number of times, the best result stored, and then GEM is used to improve the result.

In this way, we can get initializations of the plane parameters for each $\mathbf{M}_i$. The evidence can then be approximated assuming that each of the estimated plane parameters is approximately normally distributed around its mode, discounting the effect of spatial correlation. (This is not as bad as it sounds, as there is less spatial correlation between a sparse set of features, and this is only an approximation to get a rough idea as to how many layers we should use.) The details of this calculation using Laplace's approximation are not given here due to space consideration; the reader is referred to a detailed explanation given in Sivia [14, p. 88]. Fig. 1 shows the graphs of the approximated unnormalized posterior-likelihood given models $\mathbf{M}$ comprising different numbers of layers.

7.2 Initializing the Pixel Labeling

Once the number of layers has been estimated, a labeling is assigned to the high entropy points. This is done by running the GEM algorithm *just on the high entropy points* until convergence. This optimizes the segmentation of the points in which we have high confidence disparities prior to running GEM on every pixel.

8 Results

Fig. 2 shows results in two sequences of six images: the MPEG flower garden sequence and the Dayton Taylor symposium

[5] Sherlock Holmes, *The Adventure of the Sussex Vampire.*

[6] Posterior Likelihood Aggrandizement

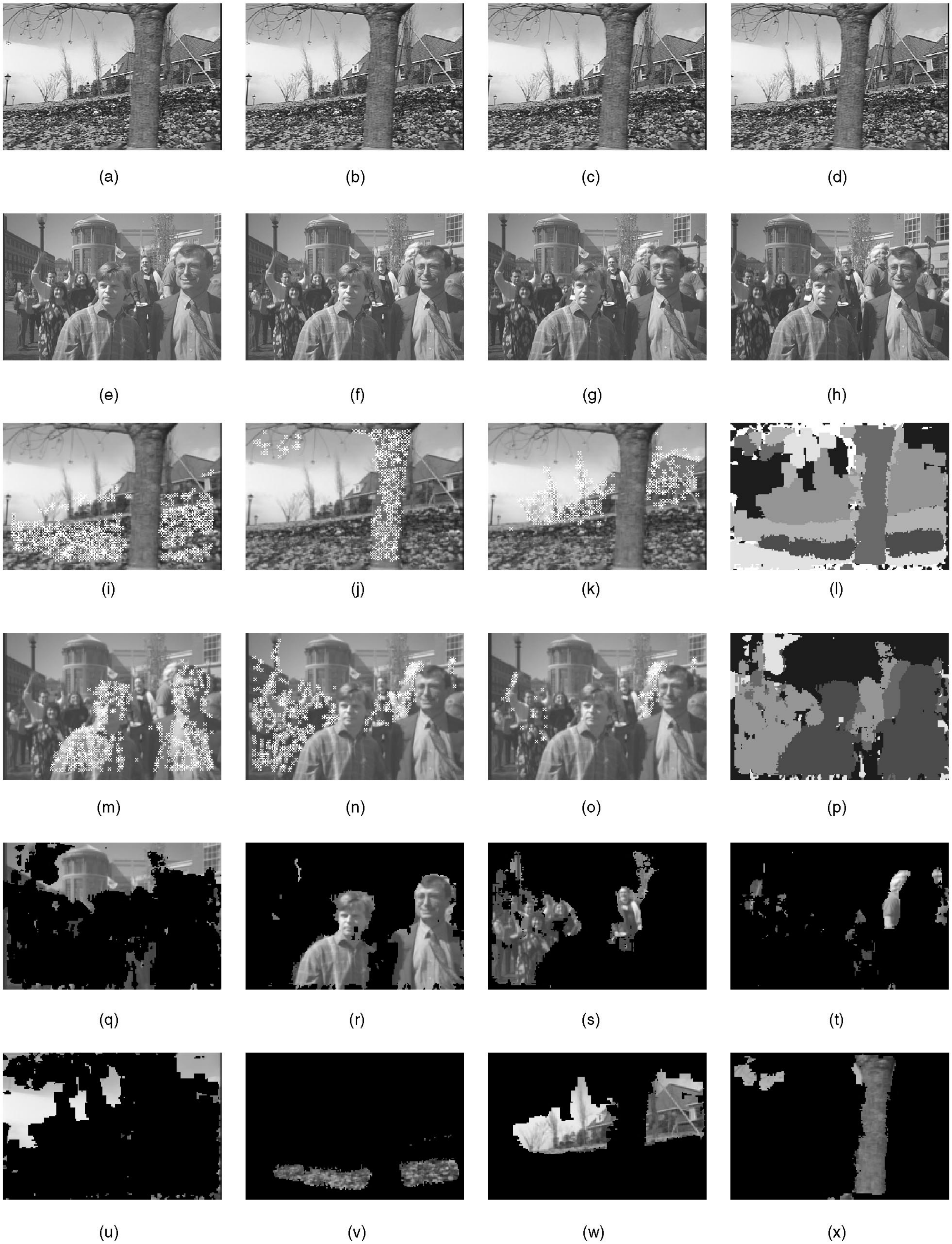

Fig. 2. (a), (b), (c), and (d) Four images of the garden sequence. (e), (f), (g), and (h) Four images of the symposium sequence (provided by kind permission of Dayton Taylor). (i), (j), and (k) Top three layers with high entropy features superimposed on them. (l) Label image, cyan represents uncertain regions that have low confidence. (m), (n), and (o) Top three layers with high entropy features superimposed on them for the Dayton Taylor sequence. (p) Label image. (q), (r), (s), (t), (u), (v), (w), and (x) Top four layers for each example sequence (note the graden is split into three segements by a horizontal ridge, see (l)).

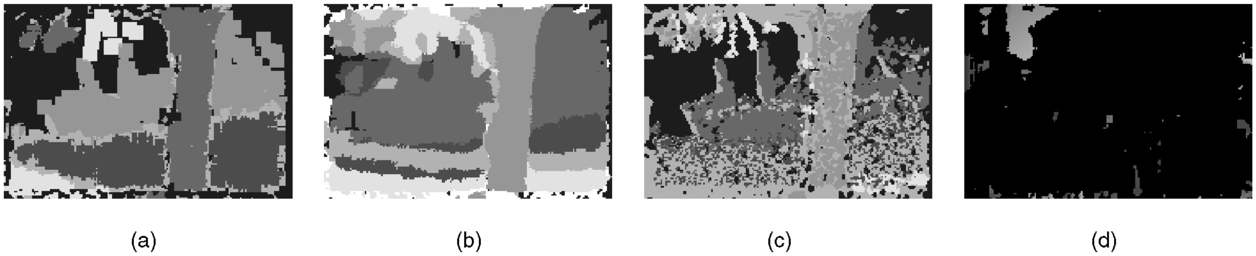

Fig. 3. (a), (b), and (c) Effects of changing the prior by large amounts: (a) increased smoothness prior, (b) fronto-parallel, (c) decreased smoothness prior (note that the twigs and branches are detected as seperate from the background), and (d) ambiguous piece of sky in the Dayton Taylor sequence.

sequence [2]. As the amount of support for the plane gets smaller, the parameters of the plane normal become increasing ill-conditioned. Note for instance that, the layer in Fig. 2o is largely determined by the white haired fellow in the red shirt and was fairly ill-conditioned. It can be likened to a door on a hinge looking for something to latch on to—in this case, it latched onto the jolly man on the left waving his arms in the air. This has happened for the layer shown in Fig. 2r to a lesser extent which is anchored by the large lady and the tree.

In the garden sequence Fig. 2v, it is very difficult to segment the sky from the trees, as in this sequence the sky provides very little motion information (being largely homogeneous), unless the prior on continuity for color is turned up. But there is always a danger in choosing our prior so that we obtain a desired result. The effects of changing the prior on the Garden sequence segmentation are shown in Fig. 3a for a prior with the smoothness (the probability two neighboring pixels have the same label controlled by β_c) turned up, Fig. 3b for a prior favoring fronto-parallel planes (lower σ_a and σ_b), and Fig. 3c with smoothness turned down. In keeping with the Bayesian approach, one should not only report the mode of the posterior but also, solutions that might be "nearly as good." Fig. 3d shows another possible layer for the Dayton Taylor sequence that could also be part of any of the other layers with little change in the posterior-likelihood as it is contains relatively little information (low entropy).

9 Summary and Conclusions

In this paper, we have developed a novel Bayesian framework for segmenting a 3D scene into plane plus parallax layers. We have demonstrated the extension of the highly successful robust estimator RANSAC to its Bayesian analogue which we have termed PILLAGE. From these, the number of planes can be estimated; the labelings of the rest of the pixels can then be initialized from the initial labeling of the high entropy points. Several pieces of evidence are aggregated within a Generalized Expectation Maximization algorithm: the original votes from the image data as to the likelihood of a given disparity, the deviation from the plane equation of a particular layer, and the spatial and color support of nearby pixels lying on the same layer.

In this paper, a zero-one utility function is assumed for choosing the number of layers. It would be interesting for specific applications to design appropriate utility functions, as this should optimize the choice of layers to take account of specific criteria. Another improvement would be to improve the priors used in this paper using ground truth (hand-labeled) training images. These are current strands of research.

A potential problem with the algorithm is that it segments purely based on the first image. This was done to get preliminary results as it was easy to implement. However, the logical progression would be to model the planar layers in 3D as in [2] and use this generative model together with the methods outlined here for automatic segmentation.

We believe that the Bayesian framework developed in this paper provides a more principled approach to estimating layers than previous approaches. Instead of relying on heuristic assumptions such as smoothness or planarity, we are able to express our prior assumptions about the scene and the imaging process explicitly. The number of layers as well as the assignment of individual pixels to layers is then automatically determined.

References

[1] S. Ayer and H. Sawhney, "Layered Representation of Motion Video," *Proc. Int'l Conf. Computer Vision,* pp. 777-784, 1995.

[2] S. Baker, R. Szeliski, and P. Anandan, "A Layered Approach to Stereo Reconstruction," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 434-441, 1998.

[3] P.N. Belhumeur, "A Bayesian Approach to Binocular Stereopsis," *Int'l J. Computer Vision,* vol. 19, no. 3, pp. 237-260, 1996.

[4] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg, "A Three-Frame Algorithm for Estimating Two-Component Image Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, no. 9, pp. 886-896, Sept. 1992.

[5] J. Besag, "Spatial Interaction and Statistical Analysis of Lattice Systems," *J. Royal Statistical Soc. London B,* vol. 36, pp. 192-225, 1974.

[6] A. Blake and A. Zisserman, *Visual Reconstruction.* Cambridge, Mass.: MIT Press, Aug. 1987.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. B,* (with discussion), vol. 39, pp. 1-38, 1977.

[8] O.D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint.* The MIT Press, 1993.

[9] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis.* Chapman & Hall, 1995.

[10] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 6, pp. 721-741, June 1984.

[11] U. Grenander, Y. Chow, and D.M. Keenan, *HANDS.* Springer-Verlag, 1991.

[12] E.T. Jaynes, "Probability Theory as Extended Logic," available at ftp://bayes.wustl.edu/pub/Jaynes/edu/pub/Jaynes/. need year.

[13] J.L. Mundy and A. Zisserman, *Geometric Invariance in Computer Vision.* MIT Press, 1992.

[14] D.S. Sivia, *Data Analysis, A Bayesian Tutorial.* Clarendon, Oxford: Clarendon, 1996.

[15] R. Szeliski, "Bayesian Modelling of Uncertainty in Low-Level Vision," *Int'l J. Computer Vision,* vol. 5, no. 3, pp. 271-301, 1990.

[16] R. Szeliski, "A Multi-View Approach to Motion and Stereo," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 157-163, 1999.

[17] J.Y.A. Wang and E.H. Adelson, "Representing Moving Images with Layers," *IEEE Trans. Image Processing,* vol. 3, no. 5, pp. 625-638, Sept. 1994.

[18] Y. Weiss and E.H. Adelson, "A Unified Mixture Framework for Motion Segmentation," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 321-326, 1996.