Parameterized Surface Fitting via MAP Estimation for Binocular Stereo

Michael J. Weisman, Alan L. Yuille and James J. Clark
Division of Applied Sciences
Harvard University
Cambridge, MA 02138

Abstract

We present a novel method for reconstructing three dimensional surfaces from stereo intensity data. We employ a set of competing surface hypotheses based on parameterized models. We use maximum a posteriori (MAP) estimation and demonstrate a connection to the Hough transform. Experimental results are given showing the effectiveness of the algorithm.

1 Introduction

Stereo vision algorithms reconstruct three dimensional surfaces from a pair of binocular images. Central to these algorithms is solving the correspondence problem. The task is to match features in one image to those in the other image. This problem is often approached through the minimization of a functional which incorporates a data term, indicating how closely features match, and a prior term, representing prior assumptions about the surface. These priors are especially important since the stereo vision problem is ill-posed in the sense that there are many possible solutions from a given pair of images. Priors are useful since they restrict the class of allowable solutions, but the danger is that if the assumptions are wrong, the prior will not produce a good reconstruction. The work described here is devoted to developing competitive priors for surface reconstructions based on a series of stereo image pairs [11]. Throughout the implementation the members of the space of priors compete. Unlikely priors are elliminated and the best (according to our goodness-of-fit criterion) remain.

If a prior is inappropriate for a specific scene, a standard algorithm will give an incorrect estimate of the depth because the assumptions about the world which give rise to that prior are not valid. For example, many smoothness assumptions commonly used induce a bias towards the frontoparallel plane [5]. As the viewpoint moves, the flattening will occur in different directions. At most one such reconstruction will be correct. This result is clearly undesirable because the reconstructed surface varies with the viewpoint. Also, parameterized models give a compact representation for incorporating new data as camera position varies.

We attempt to fit the data to a parameterized surface model. This involves simultaneously determining the model parameters and which data points lie on the surface. Eventually, we will extend this to a search over a small class of competing parameterized models [11].

2 Imaging Model and Binocular Stereo Camera Geometry

We consider the following as our basic setup. There are two cameras of known focal length f physically located at fixed points in space and directed in parallel directions. The baseline distance b is also fixed and the midpoint of the baseline connecting the two focal points is defined as the origin of our world coordinate system. The baseline defines the X-axis and the direction parallel to the lengths of the cameras determine the Z-axis. The Y-axis is orthogonal to both the X and Z axes. We define the central or 'cyclopean' camera as the virtual camera with focal point at the origin and directed parallel to the pair of cameras. A point in space (X,Y,Z) which is visible to the system of cameras projects to the left, right, and central cameras at points (x_L, y_L), (x_R, y_R), (x, y) respectively. From the geometry, it is clear that $y_L = y_R = y$. We will use the symbol y to represent this value and it should be clear from the context which y we mean. It is also clear that for any visible point projected onto the three image planes, the central camera x-coordinate will be the arithmetical average of the left and right x-coordinates. We now define disparity or parallax $u(x, y)$ as the difference in values between the coordinates x_L in the left camera and x_R in the right camera. We thus have the system of equations

$$\begin{aligned} x_L - x_R &= u \\ x_L + x_R &= 2x \end{aligned} \tag{1}$$

By similar triangles it is also readily seen that

$$\frac{x}{X} = \frac{y}{Y} = \frac{f}{Z} = \frac{u}{b} \tag{2}$$

Systems (1) and (2) are the fundamental equations of stereo vision (for parallel cameras).

IEEE International Conference
on Robotics and Automation
0-7803-1965-6/95 $4.00 ©1995 IEEE

3 Hough Energy

Here, we consider the problem of locating a surface Ξ parameterized by a set of coordinates $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)$. One example we will examine is a plane, $\Xi = \{X, Y, Z | X\cos\varphi + Y\cos\psi + Z\sin\varphi\sin\psi = \rho\}$. Here $\vec{\alpha} = (\varphi, \psi, \rho)$. The disparity $u = u(x, y, \vec{\alpha})$ is calculated from the equation for the plane and the second system of equations.

$$u = bf/Z$$
$$X = xZ/f$$
$$Y = yZ/f$$

Solving for $u(x, y, \vec{\alpha})$ yields

$$u(x, y, \vec{\alpha}) = \frac{b\cos\varphi}{\rho}x + \frac{b\cos\psi}{\rho}y + \frac{bf\sin\varphi\sin\psi}{\rho} \quad (3)$$

So for a planar surface model, the disparity is seen to be a linear function of x and y.

Given a pair of stereo images $I_L(x_L, y_L)$ and $I_R(x_R, y_R)$ we write down the following energy.

$$E[S, u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = E_1 + E_2 \quad (4)$$

where

$$E_1 = \sum_{x,y} S_{x,y}\{I_L(x + \frac{u}{2}, y) - I_R(x - \frac{u}{2}, y)\}^2$$
$$E_2 = \vartheta \sum_{x,y}(1 - S_{x,y})$$

S is a matrix represented as an image array, specifying whether a point (x, y) is on the surface Ξ or not.

$$S_{x,y} = 1 \quad (x, y) \in \Xi$$
$$S_{x,y} = 0 \quad (x, y) \notin \Xi$$

$u(x, y, \vec{\alpha})$ is defined as the disparity between the left and right images at the point (x, y) in the cyclopean coordinate system. The two terms in the energy weight the relative importance of matching intensities (or features) at each point in the central image versus having the points 'vote' for the surface. If a point does not vote for the surface, a penalty of magnitude ϑ is added to the energy. This enforces that the points remain honest for without the extra term, we would do best to have no point vote for the surface under consideration and let $S_{x,y} = 0$ everywhere.

The S field is required because all points in the visual field will only rarely lie on a single parameterized surface.

We seek a minimum for the energy and formulate the problem in the Bayesian framework via the Gibbs probability distribution. We define the probability of S and u given the surface Ξ and the data $I_L(x_L, y_L)$ and $I_R(x_R, y_R)$ as

$$P[S, u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = \frac{1}{N} e^{-\beta E[S, u(x,y,\vec{\alpha}) | I_L, I_R, \Xi]} \quad (5)$$

where N is a normalizing constant and β is a parameter inversely proportional to the 'temperature' of the system. In what follows, β will be fixed and is set based on our estimate of the noise present in the images.

The probability of the membership array S and disparity u given the surface Ξ and data $I_L(x_L, y_L)$ and $I_R(x_R, y_R)$ can be written as the product of probabilities taken over all points in the central image.

$$P[S, u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = \frac{1}{N}\prod_{x,y} e^{-\beta\{S_{x,y}(\Delta I)^2_{x,y} + \vartheta(1 - S_{x,y})\}} \quad (6)$$

where

$$(\Delta I)^2_{x,y} = \{I_L(x + \frac{u}{2}, y) - I_R(x - \frac{u}{2}, y)\}^2$$

The marginal probability of the disparity $u(x, y, \vec{\alpha})$ is obtained by summing over the space of $S_{x,y}$'s [13].

$$P_M[u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = \sum_{S_{x,y}} P[S, u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] \quad (7)$$

$$P_M[u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = \frac{1}{N}\prod_{x,y}\{e^{-\beta(\Delta I)^2_{x,y}} + e^{-\beta\vartheta}\}^1 \quad (8)$$

By multiplying over the points (x, y) in the central image, we can write the marginal probability in terms of an effective energy E_{eff} .

$$P_M[u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = \frac{1}{N}\{e^{-\beta E_{eff}[u(x,y,\vec{\alpha}) | I_L, I_R, \Xi]}\} \quad (9)$$

where

$$E_{eff}[u(x, y, \vec{\alpha}) | I_L, I_R, \Xi] = n\vartheta - \frac{1}{\beta}\sum_{x,y}\log\{1 + e^{\beta\{\vartheta - (\Delta I)^2_{x,y}\}}\} \quad (10)$$

where n is simply the number of pixels in the central image. Since n and ϑ are constant, we can drop the first term in the effective energy and write

$$H[\Xi | I_L, I_R; \vartheta, T] = T\sum_{x,y}\log\{1 + e^{\{\vartheta - (\Delta I)^2_{x,y}\}/T}\} \quad (11)$$

[1] This is a mixtures model where the first term generates points on the surface and the second term describes the unmatched points. This formulation can be seen as multiple regression (see De Veaux [4]).

where $T = \frac{1}{\beta}$ is the temperature of the system and ϑ is now seen to be a parameter controlling the allowable difference between intensities between the left and right images so that a point contributes significantly to the energy. We call $H = H[\Xi]$ the Hough energy or Hough sum for a surface Ξ due to its similarity to the standard Hough transform. In fact, in the limit as $\beta \to \infty$ (T $\to$ 0) the Hough energy approaches a Hough Transform. This relation between Hough transforms and MAP was first shown in [13]. Surfaces can now be found by finding peaks in the Hough energy.

The points with high contributions to the Hough energy and thus lie on the surface can be found as follows.

$$P[S_{x,y} = 1 | I_L, I_R, \Xi] = \frac{e^{\{\vartheta - (\Delta I)^2_{x,y}\}/T}}{1 + e^{\{\vartheta - (\Delta I)^2_{x,y}\}/T}} \tag{12}$$

We now form the image of points such that $P_{x,y} = P[S_{x,y} = 1 | I_L, I_R, \Xi]$ is greater than some threshold γ. In our implementation we take $\gamma = \frac{e}{1+e} \approx 0.73$.

4 Window Matching

4.1 Correlation

Matching straight intensities pixel-by-pixel is often not robust. Fluctuations due to camera noise and due to limitations in the pinhole camera model can cause intensities in corresponding pixels to differ. More critical are effects due to quantization and inexactitudes in the Lambertian assumption. Since the surfaces under consideration are seen from different angles except when viewing a plane front-on, the intensities received at each point will be necessarily different. A more serious problem is one of matching ambiguity among points, a point in one image will have many candidate matches in the other image. Taking a small window around each point greatly reduces the likelihood of a false match. Yang [10] argues that points which correspond in the left and right images will be locally similar. He thus argues for window matching correlations such as the sum of squared differences we employ here. To account for the possibility of one image having intensities that are scaled in the other image, we subtract the means over each window before correlating.

$$\bar{I}_L = \frac{1}{\kappa} \sum_{\xi,\eta} I_L(x_L + \xi, y_L + \eta)$$
$$\bar{I}_R = \frac{1}{\kappa} \sum_{\xi,\eta} I_R(x_R + \xi, y_R + \eta)$$

Thus, we define a feature at each point in the window as being the pixel value less the mean over a small window.

$$F_L(x_L + \xi, y_L + \eta) = I_L(x_L + \xi, y_L + \eta) - \bar{I}_L$$
$$F_R(x_R + \xi, y_R + \eta) = I_R(x_R + \xi, y_R + \eta) - \bar{I}_R$$

Now, a window correlation at each point in the central image $W(x, y)$ is defined as the sum of the feature differences squared over the window divided by κ, the number of points in the window. If the means in the left and right images differ too much, we do not allow those points to contribute to the energy.

$$W_{x,y} = \frac{1}{\kappa} \sum_{\xi,\eta} \{F_L(x_L + \xi, y_L + \eta) - F_R(x_R + \xi, y_R + \eta)\}^2 \tag{13}$$

if $|\bar{I}_L - \bar{I}_R| < \tau$ for some threshold τ, otherwise $W_{x,y} = \infty$. where x_L and x_R are given by the first system of parallel equations in terms of x and $u = u(x, y, \vec{\alpha})$.

We now write down the Hough energy in terms of the window correlation of features at each point.

$$H[\Xi | F_L, F_R; \vartheta, T] = T \sum_{x,y} \log\{1 + e^{\{\vartheta - W_{x,y}\}/T}\} \tag{14}$$

4.2 Normalized Correlation

It turns out that using a normalized correlation window matching scheme [10], the peaks in the Hough space do not change much, but the points that lie on the planar surfaces are more readily identifiable.

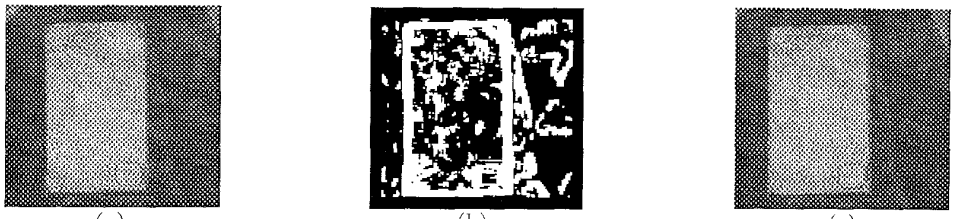
$$W_{x,y} = \frac{-\sum_{\xi,\eta} \{F_L(x_L + \xi, y_L + \eta) \cdot F_R(x_R + \xi, y_R + \eta)\}}{\sqrt{\sum_{\xi,\eta} F_L(x_L + \xi, y_L + \eta)^2} \cdot \sqrt{\sum_{\xi,\eta} F_R(x_R + \xi, y_R + \eta)^2}} \tag{15}$$

if $|\bar{I}_L - \bar{I}_R| < \tau$ for the threshold τ, otherwise $W_{x,y} = \infty$.

5 Implementation: Complications and Remedies

One serious complication with the implementation of the algorithm, is the inaccuracy of the assumption that the cameras lie on a parallel baseline. Due to the nature of the mechanical setup, it is most likely that the scan lines in one image will correspond to a scan line slightly higher or lower in the other image. We overcome this with a search over a small y-disparity at each candidate for matching pixels or matching features. Specifically, the Hough energy is modified to

$$H[\Xi | F_L, F_R; \vartheta, T] = T \sum_{x,y} \max_{\delta} \log\{1 + e^{\{\vartheta - W^{\delta}_{x,y}\}/T}\} \tag{16}$$



(a) (b) (c)

Figure 1: (a) Left image. (b) Points belonging to surface in cyclopean image. (c) Right image. Book covered with a road map in front of a sweater. The book is tilted at an angle away from the cameras. The algorithm identifies the plane with parameters $\rho = 1800\ mm$, $\varphi = \frac{11\pi}{20}$, and $\psi = \frac{3\pi}{8}$.

where

$$W^{\delta}_{x,y} = \frac{1}{\kappa}\sum_{\xi,\eta}\{F_L(x_L+\xi, y_L+\eta) - F_R(x_R+\xi, y_R+\eta+\delta)\}^2 \tag{17}$$

and δ is allowed to vary over a small range, typically $\delta \in \{0, \pm 1, \pm 2 \pm 3\}$.

6 Experimental Results

Fig. 1 shows a stereo pair of a book covered by a road map of New Jersey in front of a sweater. The book is tilted at an angle to the cameras. The algorithm is able to determine the depth of the plane, as well as the angles to the world coordinate axes within a reasonable tolerance. For this scene, the three parameters φ, ψ, and ρ give a stable and compact description. The scene viewed here is relatively simple, but by combining resulting descriptions over a more complicated scene, a compact representation is possible. A global search over the space of parameters gives a small set of possible values for the vector $\vec{\alpha}$. Peaks are found in the Hough space and are marked as possible solutions. For this plane, the MAP estimate $\rho = 1800\ mm$, $\varphi = \frac{11\pi}{20}$ and $\psi = \frac{3\pi}{8}$ yielding a maximum energy of 15,020. The constants for this run were $T = 1$, and $\vartheta = 0$. Images of 121x128 pixels were examined with normalized correlation using a 5x5 window.

7 Conclusion

We have shown a formulation for fitting stereo data to a parameterized surface model. This method involves simultaneously determining the model parameters, $\vec{\alpha}$ and which data points lie on the surface, $S = \{S_{x,y}\}$. In future work, we will extend this to a search over a small class of competing parameterized models [11]. Preliminary experiments show that a more robust reconstruction is possible by allowing a broad enough class of priors and then selecting the most appropriate one to describe parts of the scene. Competitive priors are a novel way to deal with the complications of task dependencies and viewpoint inconsistencies. The work described here demonstrated a robust procedure for fitting data to planes while simultaneously determining the set of points contributing to the plane energy and therefore lying on the planar surface. Also, our MAP estimation algorithm was shown to resemble the Hough transform in the sense that we can search over the parameter space and determine peaks corresponding to identifying surface descriptions.

Two limitations of our algorithm presented here are due to our assumptions about how data is generated in the scenes. The assumption that points either lie on a surface or are 'outliers' only remains valid when the main object covers most of the scene. The other, and perhaps more serious, issue is that the array $S = \{S_{x,y}\}$ has no bias towards spatial coherence. Thus points said to lie on the surface may turn out to be disconnected. This is possible for transparent surfaces but in general objects are made up of points locally connected. This question is addressed in recent work with Zhu [15]. A modification of the energy function as well as a unifying algorithm is discussed. Experiments lead us to be optimistic about this 'Region Competition' approach. The method also allows for multiple surfaces and models the structure of boundaries between regions.

Acknowledgment

The authors would like to thank Russell Epstein for his helpful comments and Peter Hallinan for discussions on implementation. The suggestions of the anonymous reviewers are also gratefully acknowledged.

References

[1] H. H. Baker and T. O. Binford, "Depth from edge and intensity based stereo," in *Proc. of Seventh IJCAI 1981*, Vol. 2, pp. 631–636.

[2] S. Barnard, "Stochastic stereo matching of scale," in *Int. J. Computer Vision*, Vol. 3, pp. 17–33.

[3] H. H. Bulthoff, M. Fahle, M. Wegmann, "Disparity gradients and depth scaling," *Perception*, Vol. 20, pp. 145–153.

[4] R. D. De Veaux, *Parameter Estimation for a Mixture of Linear Regression*, PhD thesis, Dept. of Statistics, Stanford University, 1986.

[5] K. Ikeuchi and B.K.P. Horn, "Numerical shape from shading and occluding boundaries," in *Artificial Intelligence*, Vol.17, No. 1-3, pp 141-184, 1981.

[6] E. T. Jaynes, "Prior Information and Ambiguity in Inverse Problems," in *SIAM-AMS Proceedings* , Vol. 14, pp 151-166, 1984.

[7] D. Jones, *Computational Models of Binocular Vision*, PhD dissertation, Dept. of Computer Science, Stanford Univ., 1991.

[8] D. Jones and J. Malik, "A Computational Framework for Determining Stereo Correspondence from a Set of Linear Spatial Filters," *Proc. of the Second ECCV*, Genova, 1992.

[9] D. Marr and T. Poggio, "A theory of human stereo vision," MIT AI Lab Memo 451, 1979.

[10] Y. Yang, *Multilevel Computation of Stereo Correspondence*, PhD thesis, Harvard University, Cambridge, MA, 1994.

[11] A. L. Yuille and J. J. Clark, "Bayesian Models, Deformable Templates, and Competitive Priors," in *Spatial Vision in Humans and Robots*, L. Harris and M. Jenkin, eds., Cambridge University Press, Cambridge, England, 1992.

[12] A. L. Yuille, D. Geiger, and Heinrich H. Bulthoff, "Stereo integration, mean field theory and psychophysics," *Network*, Vol. 2, pp. 423–442, 1991.

[13] A. L. Yuille, K. Honda and Peterson, "Particle Tracking by Deformable Templates", in *Proceedings of 1991 IEEE INNS International Joint Conference on Neural Networks,* Vol. 1, pp 7-12, Seattle, WA, 1991.

[14] R. P. Wilde, "Direct Recovery of three-dimensional scene geometry from binocular stereo disparity," *IEEE Trans. Pattern Anal. Machine Intell.*, August 1991, pp. 761–774.

[15] M. J. Weisman, A. L. Yuille, and S. C. Zhu "Region Competition for Sterco," *Harvard Robotics Laboratory Technical Report*, 1995, to appear.