MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

RECOVERING DEPTH IN STEREO CALCULATION

Jesse S. Jin, Wai-Kiang Yeap and Brian G. Cox

Department of Computer Science
University of Otago

P. O. Box 56, Dunedin
New Zealand

ABSTRACT

This paper focuses on the structure of stereo models and the techniques for measuring 3D data from a binocular visual system. A Gaussian sphere model is derived, which combines monocular cues with binocular cues by mapping a 3D space onto two 2D image planes. The determinant of the Jacobian of the mapping is given and matching is performed using zero-crossings associated with their orientation information. The possibility of transferring the knowledge such as the probability of occurrence of visual scenes to the matching process from the mapping is discussed. The triangular geometry of stereoscopic views is represented in a vector and matrix form, and the Householder transform is used in calculating depth information from stereo disparity. The distributions of error over x, y and z coordinates were analysed, providing a criterion for evaluating and comparing systems' performance.

We discuss limitations of our system: unable to cope with occlusion and transparency. These limitations are consistent with human stereopsis and in both cases we need either some high level knowledge or some other cues such as oculomotor or monocular cues to resolve the problem.

INTRODUCTION

A process which recovers 3D information from stereoscopic views has two stages, matching and stereo calculation. Different methods could be used to recover (relative) depth information from stereo and its particular choice depends on the stereo model used in the matching process. Those which employed the epipolar-line model, such as Grimson's [1981], can recover depth information from the base line and stereo disparity using an elementary geometry of triangulation. Others, like Trivedi's [1985], which allow any optical axis setting would require a singular mapping from the 3D space to two 2D images using more general vector calculations. Javis [1983] gave a detailed description about acquisition of depth information in 3D scenes. Brady [1982], and Besl and Jain [1985] provided an overview of the fields on depth information and 3D analysis.

The famous random-dot stereograms invented by Julesz have been used to show convincingly that the calculation of stereo disparity (in humans) is not based on monocularly recognizable forms such as a familiar face. The only information supplied by random-dots is the spatial position, which can cause fusion of two eyes easily in perception. Sufficient as it is, though, spatial features certainly are not the sole source for matching. Julesz [1971] gave a random-dot stereogram in which one of the images is expanded by 15%. Stereopsis can still be easily obtained, which suggests that some other information, and particularly monocular information, is important for the eyes to perform stereo matching. Another intriguing aspect of binocular vision which has long been observed is binocular rivalry [Wheatstone 1838], which refers to the alternating periods of dominance and suppression occasioned by stimulation of corresponding retinal areas with dissimilar monocular stimuli. Although there has been much empirical study of this phenomenon since then, only a few major theoretical developments have been made in stereo matching concerning binocular rivalry.

In this paper, we developed a new model which uses both monocular cues and binocular information for stereo matching. Images of two views are mapped onto a Gaussian sphere. The mapping combines monocular features, like edge and orientation, with binocular features like fixation axis and the ratio of stereo offset and focal length of cameras. To simulate binocular fusion and rivalry we implemented the model using relaxation labelling.

STEREO MODEL FOR MATCHING

In humans, the two eyes look at much the same region of visual space. Within this region of binocular overlap, the two eyes view objects from slightly different vantage points. By virtue of this lateral separation of the eyes which gives stereo disparities, humans are able to discriminate extremely small differences in relative depth. The stereo disparities include different spatial positions and orientations which can be used for matching. One problem in stereo matching using edges and orientations is that orientations and edges require different coordinate systems. The results depend critically upon the scale used to measure each coordinate. We cope with this problem by using probabilities.

Let us define the vision space as $S : X \times Y \times Z$, X, Y and Z $\subseteq R$, and consider an edge of an object passing through a point (x, y, z). If we represent this edge as an oriented

vector in 3D space, it has an angle θ with the x axis and an angle φ with the z axis. By using these two angles, the edge can also be represented as a point on the surface of a unit sphere, whose origin is (x, y, z). This is known as the Gaussian sphere [Arnold & Binford 1980], and the point is located on its surface in terms of spherical coordinates θ and φ. The Gaussian sphere defines a mapping $(\Delta x, \Delta y, \Delta z) \rightarrow (\theta, \varphi)$. Given a corresponding pair of edges, one in each image, as shown in Figure 1, we are interested in how their angles are related and how we can use this relationship to guide our matching process. Although the angles θ_l and θ_r could be of any values, they are usually of fairly similar values. This is partly due to a moderate or a small offset of the baseline.

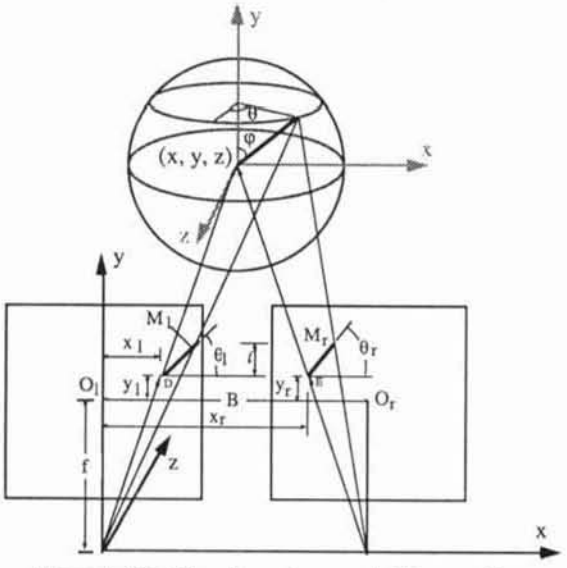

Figure 1 The Gaussian sphere model for matching

The matching process is to find the corresponding points in the left image and in the right image. We need clues to guide the search, even if the clues give only some possibilities that the features in the left image are related to the features in the right image. The orientation feature represented by the point on the Gaussian sphere casts a pair of image angles, (θ_l, θ_r), on the left image and the right image. A continuous function exists for mapping the points on the Gaussian sphere, with coordinates θ and φ, to the image angles (θ_l, θ_r), i.e. $\theta \times \varphi \rightarrow \theta_l \times \theta_r$, . Similarly, there is an inverse function P which maps points in the space $\theta_l \times \theta_r$ to points on the Gaussian sphere, $\theta \times \varphi$. From the probability theorem, the probability distribution of (θ_l, θ_r) equals the probability distribution of (θ, φ) multiplying with the Jacobian determinant of the mapping P [Blake 1979]. If we suppose all edges of objects are randomly and uniformly distributed in the (θ, φ) domain, the probability distribution ψ of (θ_l, θ_r) will be

$$\psi(\theta_l, \theta_r) = \frac{1}{A} |J_P|$$

where A is the area of the definition domain Ω of (θ, φ).

This distribution gives a correlation function for θ_l and θ_r.

The basic steps of our stereo process can now be stated as:

1. build a geometric model of binocular visual system, based on the Gaussian sphere,
2. find a mapping P: $\theta_l \times \theta_r \rightarrow \theta \times \varphi$,
3. calculate the determinant of the Jacobian of P
$$\det J_P = \left| \frac{\partial(\theta, \varphi)}{\partial(\theta_l, \theta_r)} \right| = \frac{\partial\theta}{\partial\theta_l}\frac{\partial\varphi}{\partial\theta_r} - \frac{\partial\theta}{\partial\theta_r}\frac{\partial\varphi}{\partial\theta_l}$$
4. suppose (θ, φ) are in a uniform distribution and calculate the distribution function of $\psi(\theta_l, \theta_r)$,
5. extract features $(\theta_{l_1}, \theta_{l_2}, \cdots, \theta_{l_m})$ from the left image and features $(\theta_{r_1}, \theta_{r_2}, \cdots, \theta_{r_n})$ from the right image,
6. perform relaxation labelling on θ_{l_i} and θ_{r_j} to get an optimal matching.

The determinant of the Jacobian of the mapping P: $(\theta_l, \theta_r) \rightarrow (\theta, \varphi)$ is derived by

1. find a mapping Q: $\theta \times \varphi \rightarrow \theta_l \times \theta_r$, with coordinates (x, y, z),
2. inverse the mapping Q to obtain P by:
   2.1 inverse Q under the same coordinates (x, y, z),
   2.2 transform (x, y, z) coordinates into (θ, φ) coordinates, giving the mapping
   P: $\theta_l \times \theta_r \rightarrow \theta \times \varphi$,
3. calculate the Jacobian matrix J_P of the mapping P,
4. calculate and simplify the determinant $|J_P|$.

When the visual distance z is far enough comparing with B, i.e. $B/z \ll 1$, we have the determinant of the Jacobian matrix defined as:

$$|J_P| \approx \frac{y_l [(x_r - x_l) - B\cos^2\theta_l]}{(x_l^2 + y_l^2 + 1)\sqrt{x_l^2 + y_l^2}\,\sin^2(\theta_l - \theta_r)} \qquad (1)$$

The detailed deduction can be found in [Jin 1992].

It is noteworthy that the mapping P is not a bijective mapping. It is not defined at (0, 0), as the circle $z = 0$ of points on the sphere for which $\theta = 0$ all map to (0, 0). The mapping is not invertible at that point, which is why we use P to represent the mapping $\theta_l \times \theta_r \rightarrow \theta \times \varphi$ rather than Q^{-1}. This fact tallies with the effect in human vision. When people view a horizontal wire, they often lose their depth perception. This is because the uniform texture on the wire wipes out the size perception so that the stereo matching depends solely on orientation, but the zero orientations in both eyes fail to stimulate binocular neurons to cause fusion.

The value of $r = \sqrt{x_l^2 + y_l^2}$ in formula (1) represents the eccentricity of the Jacobian determinant. Mapping factors reduce eccentrically which tallies with the fact that visual acuity decreases with retinal eccentricity.

DEPTH CALCULATION

In Trivedi's [1985] model of stereopsis, a stereoscopic

view reflects the geometry of two cameras, as shown in Figure 2. The stereoscopic area (the dotted area) is the overlap between the fields of view of the two cameras.

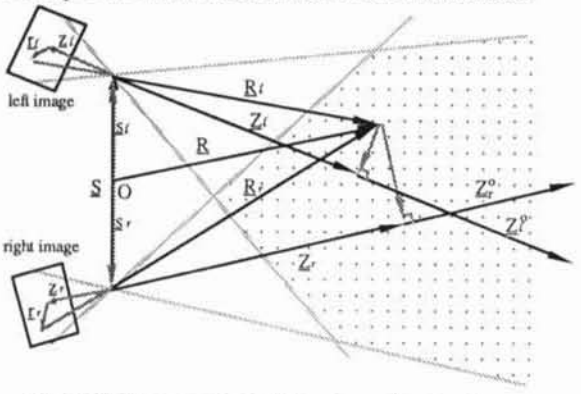

Figure 2 The geometrical structure of a visual system

In three dimensions the supporting matched pairs corresponding to physical points provide a viewer-centred scene description called the geometrical descriptive base [Pollard et al. 1987]. In this base, each point is represented by a vector R. We shall use an underlined letter to represent a directional vector, e.g., $\underline{R}$, and use a small superscript zero to mark a unit directional vector, e.g., $\underline{R}^*$. We shall use upper case letters to refer to the scene in three dimensions while lower case letters refer to the images. Subscripts "l" and "r" refer to the left view and the right view, respectively. The geometrical origin O is chosen in the middle of two foci of the left and the right cameras as shown in Figure 2 without loss of generality.

Considering the four triangles bounded by the pairs of vectors $(\underline{r}_l, z_l)$, $(\underline{r}_r, z_r)$, $(\underline{R}_l, \underline{Z}_l)$ and $(\underline{R}_r, \underline{Z}_r)$, we have

$$\underline{r}_l/z_l = (\underline{Z}_l - \underline{R}_l)/Z_l = \underline{Z}_l^* - (\underline{R} - \underline{s}_l)/Z_l \qquad (2)$$

and

$$\underline{r}_r/z_r = (\underline{Z}_r - \underline{R}_r)/Z_r = \underline{Z}_r^* - (\underline{R} - \underline{s}_r)/Z_r \qquad (3)$$

According to the initial setting, we have $\underline{s} = \underline{s}_l - \underline{s}_r$ i.e.

$$\underline{s} = \underline{s}_l - \underline{s}_r = (\underline{r}_l/z_l - \underline{Z}_l^*)Z_l - (\underline{r}_r/z_r - \underline{Z}_r^*)Z_r \qquad (4)$$

and

$$2\underline{R} = -[(\underline{r}_l/z_l - \underline{Z}_l^*)Z_l + (\underline{r}_r/z_r - \underline{Z}_r^*)Z_r] \qquad (5)$$

Formula (4) is a vector equality in three dimensions. It can be rewritten as

$$\begin{aligned} s_x &= (r_{lx}/z_l - Z_{lx}^*)Z_l - (r_{rx}/z_r - Z_{rx}^*)Z_r \\ s_y &= (r_{ly}/z_l - Z_{ly}^*)Z_l - (r_{ry}/z_r - Z_{ry}^*)Z_r \\ s_z &= (r_{lz}/z_l - Z_{lz}^*)Z_l - (r_{rz}/z_r - Z_{rz}^*)Z_r \end{aligned} \qquad (6)$$

These are consistent equations, and Z_l and Z_r can be obtained by solving these equations. The solution of Z_l and Z_r from (6) can be used in (5) to calculate $\underline{R}$, which gives the 3D coordinates of the point in the space of the scene.

Given the above mathematical model, it is not difficult to calculate the required depth information. In practice, however, one faces numerous problems due to imprecise data. To solve this problem, we introduce an efficient algorithm using the Householder transform [Strang 1988]. Two Householder transforms H_2 and H_1 transform A into an upper triangular matrix, and $H_2H_1AZ = H_2H_1S$ can be solved by a back substitution. Replacing the solution Z in (5), we have

$$R = \begin{pmatrix} R_x \\ R_y \\ R_z \end{pmatrix} = \frac{1}{2}\begin{pmatrix} (Z_{lx}^o - r_{lx}/z_l)Z_l + (Z_{rx}^o - r_{rx}/z_r)Z_r \\ (Z_{ly}^o - r_{ly}/z_l)Z_l + (Z_{ry}^o - r_{ry}/z_r)Z_r \\ (Z_{lz}^o - r_{lz}/z_l)Z_l + (Z_{rz}^o - r_{rz}/z_r)Z_r \end{pmatrix} \qquad (7)$$

The derivation of the algorithm using the Householder transform can be found in [Jin 1992].

THE DISTRIBUTION OF ERRORS

One of the fundamental problems in stereo vision is that the accuracy of the measurements decreases as the distance increases. In order to find out the distribution of stereo disparity, we simplify the arrangement of the stereo model. There is no loss of generality in assuming that the two image screens are coplanar, and the two x axes parallel to each other. If we take the focal length f of the two cameras as a standard unit, we have $z_l = z_r = 1$, $Z = Z_l = Z_r$ and $\underline{Z}_l^* = \underline{Z}_r^*$. We define $\lambda = |\underline{S}|/Z$, which gives a parameter to analyse the relation between the distance and the accuracy of the measurements.

Defining $\delta = |(\underline{r}_l - \underline{r}_r)|$, which is stereo disparity, from the assumption and formula (4), we have $\underline{s} = (\underline{r}_l - \underline{r}_r)Z$, i.e., $\underline{s}/Z = (\underline{r}_l - \underline{r}_r)$. From the definitions of λ and δ,we have

$$\lambda = |\underline{s}|/Z = |(\underline{r}_l - \underline{r}_r)| = \delta \qquad (8)$$

Formula (8) indicates several facts:

First, $\lambda = \delta$ indicates that a large λ gives a large stereo disparity. Their differentiation $\frac{d\lambda}{d\delta} = 1$ shows that stereo disparity increases with increasing λ.

Second, rearranging (8) we have $|\underline{s}| = \delta Z$. To calibrate, suppose there is a setup error differentiating $|\underline{s}|$ over Z, we have $\frac{d|\underline{s}|}{dZ} = \delta \geq 0$, which means that the accuracy of Z increases with the value of offset $|\underline{s}|$.

Third, for measuring we have $\delta = |\underline{s}|/Z$. For a setup system, i.e. $|\underline{s}|$ is a constant, differentiating δ over Z, we have $\frac{d\delta}{dZ} = -\frac{|\underline{s}|}{Z^2}$, which means the measurement error is inversely related to the square of the distance.

The error distributions over x and y coordinates are analysed as follows. From formula (7) and $|\underline{R}| = \sqrt{(R_x^2 + R_y^2 + R_z^2)}$, partially differentiating $|\underline{R}|$ and noting that Z is independent of x and y by definition, we have

$$\frac{\partial|R|}{\partial x} = \frac{R_x}{\sqrt{R_x^2 + R_y^2 + R_z^2}}\frac{\partial R_x}{\partial x} = \frac{1}{|R|}(r_{lx} + r_{rx})Z^2\frac{\partial(r_{lx}+r_{rx})}{\partial x} \qquad (9)$$

Formula (9) gives the relation between the difference of measurements over the x coordinate $\frac{\partial|R|}{\partial x}$ and the error in the original data $\frac{\partial(r_{lx}+r_{rx})}{\partial x}$. Rewriting (9), we have

$$\frac{\partial |R|}{\partial x} = \frac{R_x}{\sqrt{R_x^2 + R_y^2 + R_z^2}} \frac{\partial R_x}{\partial x} = \frac{1}{|R|}[(r_{fx} + s_x/2) + (r_{rx} - s_x/2)]Z^2 \frac{\partial (r_{fx} + r_{rx})}{\partial x}$$

where $(r_{fx} + s_x/2)$ and $(r_{rx} - s_x/2)$ are focusing centres on the x axis of the left and right views respectively. From this we can draw an error distribution as shown in Figure 3, where the curves give the equal-error contours.

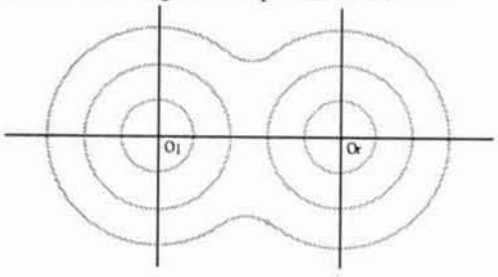

Figure 3 The error distribution over the x coordinate

Similarly, the relation over the y coordinate is

$$\frac{\partial |R|}{\partial y} = \frac{R_y}{\sqrt{R_x^2 + R_y^2 + R_z^2}} \frac{\partial R_y}{\partial y} = \frac{1}{|R|}(r_{fy} + r_{ry})Z^2 \frac{\partial (r_{fy} + r_{ry})}{\partial y}$$

and its error distribution is shown in Figure 4.

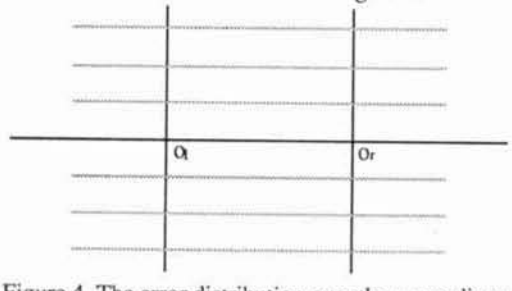

Figure 4 The error distribution over the y coordinate

DISCUSSION AND CONCLUSION

The significance of the mapping is that it defines a relation between the visual world and the two stereoscopic views, and combines the monocular features from two views with the binocular correlation of the stereopsis. The mapping as defined allows us to manipulate the model in various ways and reflect several characteristics of stereo vision in humans. First, any a priori knowledge about the world, either from our knowledge of the visual scenes or from the features extracted monocularly from the stereoscopic views, can be applied in the mapping. Second, we can adjust the focus of the view point either to improve the success rate of the stereo matching or the increased accuracy of the stereo disparity. The distribution of the determinant of the Jacobian varies with $x_l^2+y_l^2$ (i.e. concentrically). Close to the centre, we have a steep distribution along $\theta_l = \theta_r$ which gives more weight for matching the stereo than that for calculating stereo disparity, and vice versa.

Our solution to equation (6) is insensitive to errors as a result of the smoothing effect in the least squares approximation. The Household transform is an efficient method for solving equation (6) from the viewpoint of both space and time.

Increasing the baseline of the two cameras will help improve the accuracy of the measurements but this will also increase the difficulty of stereo matching since more widely differing views of the object can be obtained. A solution reducing the problem in the matching process is preferred since it is the more crucial part of the stereo problem.

REFERENCES

Arnold, R D & Binford, T O (1980). Geometric constraints in stereo vision. *Proc. SPIE: Image Processing for Missile Guidance* 238:281-292.

Besl, P J & Jain, R C (1985). Three-dimensional object recognition. *Computing Surveys* 17:75-145.

Blake, I F (1979). *An introduction to applied probability*. NY: John Wiley & Sons.

Brady, M (1982). Computational approaches to image understanding. *Computing Surveys* 14:3-71.

Grimson, W E L (1981). *From Images to Surfaces*. Cambridge: MIT Press.

Jarvis, R A (1983). A perspective on range finding techniques for computer vision. *IEEE Trans. on PAMI* 5:122-139.

Jin, J S (1992). Depth Acquisition and Surface Reconstruction in Three-dimensional Computer Vision. Ph.D. Thesis, Dunedin: University of Otago.

Julesz, B (1971). *Foundation of cyclopean perception*. Chicago: University of Chicago Press.

Pollard, S B; Porrill, J; Mayhew, J E W & Frisby, J P (1987). Matching geometrical descriptions in three-space. *Image and Vision Computing* 5:73-78.

Strang, G (1988). *Linear Algebra and Its Applications*. San Diego: Harcourt Brace Jovanovich, Inc.

Trivedi, H P (1985). A computation theory of stereo vision. In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, CA, 277-282.

Wheatstone, C (1838). Contributions to the physiology of vision: 1. On some remarkable and hitherto unobserved phenomena of binocular vision. *Proc. of the Royal Society of London* B18:371-394.