Stereo Without Disparity Gradient Smoothing: a Bayesian Sensor Fusion Solution

Ingemar J. Cox, Sunita Hingorani,
Bruce M. Maggs and Satish B. Rao

NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, U.S.A.

Abstract

A maximum likelihood stereo algorithm is presented that avoids the need for smoothing based on disparity gradients, provided that the common uniqueness and monotonic ordering constraints are applied. A dynamic programming algorithm allows matching of the two epipolar lines of length N and M respectively in $O(NM)$ time and in $O(N)$ time if a disparity limit is set. The stereo algorithm is independent of the matching primitives. A high percentage of correct matches and little smearing of depth discontinuities is obtained based on matching individual pixel intensities. Because feature extraction and windowing are unnecessary, a very fast implementation is possible.

Experiments reveal that multiple global minima can exist. The dynamic programming algorithm is guaranteed to find one, but not necessarily the same one for each epipolar scanline. Consequently, there may be small local differences between neighboring scanlines.

1 Introduction

Stereo algorithms seek to find corresponding features between a *pair* of images. Stereo algorithms can be characterized by (1) the primitive features that are matched, (2) the *local* cost of matching two features and (3) the *global* cost function and associated constraints. The stereo framework presented here is, at the algorithmic level, independent of the feature primitives. However, for the experimental results of Section (3), matching was performed directly on the scalar intensity values of the individual pixels. Matching occurs along epipolar lines which are assumed, for convenience, to be coplanar with the image scanlines. The epipolar constraint reduces the stereo correspondence problem from two to one dimension. Most, if not all, previous stereo algorithms include a cost based on the disparity gradient [1, 2, 4, 5, 11], i.e., the difference in depth between two pixels divided by their distance apart. This cost can be

BMVC 1992 doi:10.5244/C.6.35

thought of as a regularization factor [10] which serves to constrain surfaces to be smooth. However, surfaces are not smooth at depth discontinuities which are the most important features of depth maps. One contribution of this paper is to show that penalizing disparity gradients is unnecessary, provided that the common assumptions of uniqueness and monotonic ordering are made. This is detailed in Section (2), in which stereo is formulated as a Bayesian sensor fusion problem. A local cost function is derived that does not penalize disparity gradients. Section (2.1) then describes how a global minima can be found using a dynamic programming algorithm that enforces the uniqueness and monotonicity constraints. The experiments described in Section (3) reveal that multiple *global* minimum may exist. This can give rise to (minor) artifacts in the disparity map. Similar multiple global minima may exist for other stereo algorithms. Results for several natural scenes are included. Finally, Section (4) concludes with a discussion of the advantages and disadvantages of this algorithm and possible future work.

2 Deriving Cost Functions

In this section, the cost of matching two features, or declaring a feature occluded is first derived, then a global cost function that must be minimized is derived. To begin, we introduce some terminology as developed by Pattipati *et al* [9]. Let the two cameras be denoted by $s = \{1, 2\}$ and let $\mathbf{Z}_s$ represent the set of measurements obtained by each camera along corresponding epipolar lines: $\mathbf{Z}_s = \{\mathbf{z}_{s,i_s}\}_{i_s=0}^{m_s}$ where m_s is the number of measurements from camera s and $\mathbf{z}_{s,0}$ is a dummy measurement, the matching to which indicates no corresponding point. For epipolar alignment of the scanlines, $\mathbf{Z}_s$ is the set of measurements along a scanline of camera s. The measurements $\mathbf{z}_{s,i_s}$ might be simple scalar intensity values or higher level features. Each measurement $\mathbf{z}_{s,i_s}$ is assumed to be corrupted by additive, white noise.

The condition that measurement $\mathbf{z}_{1,i_1}$ from camera 1, and measurement $\mathbf{z}_{2,i_2}$ from camera 2 originate from the same location, x_k, in space, i.e. that $\mathbf{z}_{1,i_1}$ and $\mathbf{z}_{2,i_2}$ correspond to each other is denoted by Z_{i_1,i_2}. The condition in which measurement $\mathbf{z}_{1,i_1}$ from camera 1 has no corresponding measurement in camera 2 is denoted by $Z_{i_1,0}$ and similarly for measurements in camera 2. Thus, $Z_{i_1,0}$ denotes occlusion of feature z_{1,i_1} in camera 2.

Next, we need to calculate the *local* cost of matching two points $\mathbf{z}_{1,i_1}$ and $\mathbf{z}_{2,i_2}$. The likelihood that the measurement pair Z_{i_1,i_2} originated from the same point x_k is denoted by $\Lambda(Z_{i_1,i_2} \mid x_k)$ and is given by

$$\Lambda(Z_{i_1,i_2} \mid x_k) = \prod_{s=1}^{2} [P_{D_s} p(\mathbf{z}_{s,i_s} \mid x_k)]^{1-\delta_{i_s}} [1 - P_{D_s}]^{\delta_{i_s}} \tag{1}$$

where δ_{i_s} is an indicator variable that is unity if a measurement is not assigned a corresponding point, i.e. is occluded, and zero otherwise. The term $p(\mathbf{z} \mid x)$ is a probability density distribution that represents the likelihood of measurement $\mathbf{z}$ assuming it originated from a point x in the scene. The parameter P_{D_s} represents the probability of detecting a measurement originating from x_k at sensor s. This parameter is a function of the number of occlusions, noise etc. Conversely, $(1 - P_D)$ may be viewed as the probability of occlusion. If it is assumed that the measurements vectors $\mathbf{z}_{s,i_s}$ are normally distributed about their ideal value $\mathbf{z}$, then

$$p(\mathbf{z}_{s,i_s} \mid x_k) = \mid (2\pi)^d \mathbf{S}_s \mid^{-\frac{1}{2}} exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_{s,i_s})' \mathbf{S}_s^{-1} (\mathbf{z} - \mathbf{z}_{s,i_s}) \right\} \tag{2}$$

where d is the dimension of the measurement vectors $\mathbf{z}_{s,i_s}$ and $\mathbf{S}_s$ is the covariance martix associated with the error $(\mathbf{z} - \mathbf{z}_{s,i_s})$. Since the true value, $\mathbf{z}$, is unknown we approximate it by maximum likelihood estimate $\hat{\mathbf{z}}$ obtained from the measurement pair Z_{i_1,i_2} and given by

$$\mathbf{z} \approx \hat{\mathbf{z}} = \mathbf{S}_{2,i_2}(\mathbf{S}_{1,i_1} + \mathbf{S}_2)^{-1} \mathbf{z}_{1,i_1} + \mathbf{S}_{1,i_1}(\mathbf{S}_{1,i_1} + \mathbf{S}_{2,i_2})^{-1} \mathbf{z}_{2,i_2} \tag{3}$$

where $\mathbf{S}_{s,i_s}$ is the covariance associated with measurement $\mathbf{z}_{s,i_s}$.

Now that we have established the cost of the individual pairings Z_{i_1,i_2}, it is necessary to determine the total cost of all pairs. Denote by γ a feasible pairing of all measurements and let Γ be the set of all feasible partitions, i.e. $\Gamma = \{\gamma\}$. If γ_0 denotes the case where all measurements are unmatched, i.e., the case in which there are no corresponding points in the left and right images, then we wish to find the pairings or partition γ that maximizes $L(\gamma)/L(\gamma_0)$ where the likelihood $L(\gamma)$ of a partition is defined as

$$L(\gamma) = p(\mathbf{Z}_1, \mathbf{Z}_2 \mid \gamma) = \prod_{Z_{i_1,i_2} \epsilon \gamma} \Lambda(Z_{i_1,i_2} \mid x) \left(\frac{1}{\phi_1}\right)^{n_1} \left(\frac{1}{\phi_2}\right)^{n_2} \tag{4}$$

where ϕ_s is the field of view of camera s and n_s is the number of unmatched measurements from camera s in partition γ. The likelihood of no matches, $L(\gamma_0)$ is therefore given by $L(\gamma_0) = 1/(\phi_1^{n_1}\phi_2^{n_2})$

The maximization of $L(\gamma)/L(\gamma_0)$ is eqivalent to

$$\min_{\gamma\epsilon\Gamma} J(\gamma) = \min_{\gamma\epsilon\Gamma} [\ln(L(\gamma_0)) - \ln(\mathrm{L}(\gamma))] \tag{5}$$

which leads to

$$\begin{aligned}\min_{\gamma\epsilon\Gamma} J(\gamma) &= \min_{\gamma\epsilon\Gamma} \sum_{Z_{i_1,i_2}\epsilon\gamma} \left\{ \sum_{s=1}^{2} \left\{ (1-\delta_{i_s}) \left[\frac{1}{2}(\hat{\mathbf{z}} - \mathbf{z}_s)'\mathbf{S}_s^{-1}(\hat{\mathbf{z}} - \mathbf{z}_s) \right] + \right.\right. \\ &\left.\left. \delta_{i_s} \left[\ln\left(\frac{P_{D_s}}{1 - P_{D_s}} \frac{1}{\mid (2\pi)^d \mathrm{S}_s^{-1} \mid^{\frac{1}{2}}} \right) \right] \right\} \right\} \end{aligned} \tag{6}$$

The first term in the inner summation of Equation (6) is the cost of matching two features while the second term is the cost of an occlusion/disparity discontinuity. Clearly, as the probability of occlusion $(1 - P_{D_s})$ becomes small the cost of not matching a feature increases, as expected.

2.1 Dynamic Programming Solution

The minimization of Equation (6) is a classical weighted matching or assignment problem [8]. There exist well known algorithms for solving this with polynomial complexity $O(N^3)$ [7]. If the assignment problem is applied to the stereo matching problem directly, non-physical solutions are obtained. This is because Equation (6) does not constrain a match at $\mathbf{z}_{i_s}$ to be close to the match for $\mathbf{z}_{(i-1)_s}$, yet surfaces are usually smooth, except at depth discontinuities. In order to impose this smoothness condition, previous researchers have included a disparity gradient term to their cost function [1, 4, 5, 11, 12]. The problem with this approach is that it tends to blur the depth discontinuities as well as introduce additional free parameters that must be adjusted.

Instead, we assume as in [6] (1) *uniqueness*, i.e. a feature in the left image can match to no more than one feature in the right image and vice versa and (2) monotonic ordering, i.e. if $\mathbf{z}_{i_1}$ is matched to $\mathbf{z}_{i_2}$ then the subsequent measurement $\mathbf{z}_{i_1+1}$ may only match measurements $\mathbf{z}_{i_2+j}$ for which $j > 0$. The minimization of Equation (6) subject to these constraints can be solved by dynamic programming. If there are N and M measurements in each of the two epipolar scanlines, respectively, then Ohta and Kanade [6] presented a solution with complexity $O(N^2M^2)$. We have improved this minimization procedure to $O(NM)$:

```
Occlusion = $\left[\ln\left(\frac{P_{D_s}}{1-P_{D_s}}\frac{1}{|(2\pi)^d \mathbf{S}_s^{-1}|^{\frac{1}{2}}}\right)\right]$
for (i=1;i$\leq$ N;i++){ C(i,0) = i*Occlusion }
for (i=1;i$\leq$ M;i++) { C(0,i) = i*Occlusion}
for(i=1;i$\leq$ N;i++){
   for(j=1;j$\leq$ M;j++){
      C(i,j) = min (C(i-1,j-1)+c($z_{1,i}$,$z_{2,j}$), C(i,j-1)+Occlusion,
                    C(i-1,j)+Occlusion) } }
```

where $C(i, j)$ represents the cost of matching the first i features in the left image with the first j features in the right image and $c(\mathbf{z}_{1,i}, \mathbf{z}_{2,j})$ is the cost of matching the two features $\mathbf{z}_{1,i}, \mathbf{z}_{2,j}$ as shown in Equation (6).

Of course, this general solution can be further improved by realizing that there is a practical limit to the disparity between two measurements. This is also true for human stereo, the region of allowable disparity being referred to as Panum's fusional area [3]. If a measurement $\mathbf{z}_{i_1}$ is constrained to match only measurements $\mathbf{z}_{i_2}$ for which $i_1 - \Delta x \leq i_2 \leq i_1 + \Delta x$ then the time required by dynamic programming algorithm can be reduced to linear complexity $O(N)$.

3 Experimental Results

Unless otherwise stated, all experiments described here were performed with scalar measurement vectors representing the intensity values of the individual pixels, i.e. $\mathbf{z}_{i_s} = I_{i_s}$. The field of view of each camera, ϕ_s, is assumed to be π and the measurements are assumed to be corrupted with white noise of variance $\sigma^2 = 16$. Finally, the probability of detection P_{D_s} is assumed to be 0.9 so that the cost of an occlusion is 3.8.

3.1 Random Dot Stereograms

Figure (1) shows the depth map obtained from the left image of a "wedding cake" random dot stereogram - three rectangular regions one above the other. Note that black pixel values indicate no match with pixels in the right image. While the number of correct matches is 95.4%, it is interesting to examine why the correct depth estimates have not been found at every point on every line. In particular, since the RDS pair is noise free, a perfect match is expected, so the right side of each rectangle should exhibit a depth discontinuity that is aligned with neighboring scanlines. This is not the case in practice. Close examination of this phenomenon revealed there are multiple *global* minima! Dynamic programming is guaranteed to find a global minima but not necessarily the same one for each scanline. Hence, the misalignment of the vertical depth discontinuities. This is a problem. Note however, that the jagged vertical discontinuity caused by the multiple global minima is a characteristic of other stereo algorithms [2, 6] and may be indicative of the presence of multiple global minima in other stereo algorithms.

Rather than choose an arbitrary solution from amongst the set of global minimum, a second optimization can be performed that selects from the set of solutions, that solution which contains the least number of discontinuities. Performing this minimization *after* first finding all maximum likelihood solutions is very different from incorporating the discontinuity penalty into the original cost. The second level of minimization can be easily accomplished as part of the dynamic programming algorithm without having to enumerate all maximum likelihood solutions.. The result of applying the maximum likelihood minimum discontinuity algorithm to the random dot stereogram is shown in Figure (2). A significant improvement is evident, with the percentage of correct matches increasing to 98.7%. Once again, multiple global minima are evident but their number is far fewer.

Note that using the dynamic programming algorithm with a disparity limit of 25 pixels a 256x256 pixel image pixel scanline takes approximately 11 seconds on a SGI Personal Iris. Each scanline therefore takes 0.04 seconds which is very close to video rates of 0.033 seconds per frame, if all scanlines are processed in parallel.

3.2 Natural Scenes

Figure (3) is the left image of the "Pentagon" stereogram. Figures (4) and (5) shows the resulting disparity maps for the maximum likelihood (ML) and ML with minimum discontinuities (MLMD) algorithms. The MLMD provides a qualitative improvement. Note that for display purposes, those pixels that were not matched are assigned the disparity value of whichever of the left or right neigboring pixel is furthest away. Once again, vertical depth discontinuities exhibit some misalignment between scanlines. Nevertheless, significant detail is obtained, as is evident from the overpasses and freeways in the upper right corner of the image. Figure (6) shows the result of applying the MLMD algorithm for $P_D = 0.99$ and supports our observation that the algorithm is stable for reasonable variations in the free parameter value.

Figures (8) and (9) show the results of applying the ML and MLMD algorithm to a stereo pair, the left image of which is shown in Figure (7). Especially noteworthy is the narrow sign pole in the middle right of the image which illustrates the sharp depth detail that is extracted.

The algorithm was tested on other stereograms and similar performance was obtained. However page restrictions, prevent further examples.

4 Conclusion

Determining the correspondence between two stereo images was formulated as a Bayesian sensor fusion problem. A local cost function was derived that consists of (1) a normalized squared error term that represents the cost of matching two features and (2) a fixed penalty for an unmatched measurement that is a function of the probability of occlusion. These two terms are common to other stereo algorithms, but the additional smoothing term based on disparity gradients is avoided. Instead, uniqueness and monotonicity constraints, imposed via a dynamic programming algorithm constrain the solution to be physically sensible.

The dynamic programming algorithm has complexity $O(NM)$ which reduces to $O(N)$ if a disparity limit is set. The algorithm is potentially very fast. especially since a high percentage of correct matches were obtained on intensity based matching primitives that require no feature extraction.

Experimental results were presented for RDS and natural images with good results. The random dot stereograms revealed that multiple *global* minima may exist. Consequently, there may be small local differences between neighboring scanlines. Similar differences are visible for other stereo algorithms which may indicate that multiple global minima are a problem for these algorithms as well. A more detailed study of this phenomenon is needed. In particular, does a sensible cost function with only a single global minima exist?

The experimental results described here do *not* use any information between scanlines. This is somewhat surprising, but was a concious decision to avoid blurring horizontal depth discontinuities. The maximum likelihood minimum horizontal discontinuities (MLMD) also suffers from multple global minima, though far fewer than the maximum likelihood algorithm alone. A third level of optimization should be investigated that maximizes the continuity between scanlines. This is being examined.

Acknowledgements

Thanks to Yaakov Bar-Shalom and Davi Geiger for valuable discussion on issues related to this paper. Also thanks to Takeo Kanade and Tomoharu Nakahara for supplying several stereo images. Special thanks to K.G. Lim of Cambridge University for the interest shown in this algorithm. Thanks to George V. Paul for implementation assistance.

References

[1] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, 1987.

[2] D. Geiger, B. Ladendorf, and A. L. Yuille. Binocular stereo with occlusions. In *Second European Conference on Computer Vision*, 1992.

[3] D. Marr. *Vision.* W. H. Freeman & Co., 1982.

[4] D. Marr and T. Poggio. A cooperative stereo algorithm. *Science*, 194, 1976.

[5] J. E. W. Mayhew and J. P. Frisby. Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, 17, 1981.

[6] Y. Ohta and T. Kanade. Stereo by intra- and inter- scanline search using dynamic programming. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-7(2):139–154, 1985.

[7] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization.* Prentice Hall, 1982.

[8] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity.* Prentice Hall, 1982.

[9] K. R. Pattipati, S. Deb, and Y. Bar-Shalom. Passive multisensor data association using a new relaxation algorithm. In *Multitarget-Multisensor Tracking: Advanced Applications*, pages 219–246. Artech House, 1990.

[10] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:638–643, 1985.

[11] K. Prazdny. Detection of binocular disparities. *Biological Cybernetics*, 52, 1985.

[12] A. L. Yuille, D. Geiger, and H. Bulthoff. Stereo integration, mean field theory and psychophysics. In *First European Conference on Computer Vision*, pages 73–82, 1990.

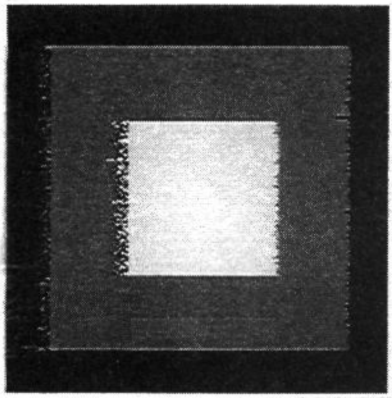

Fig 1: Maximum likelihood disparity map for random dot stereogram with $P_D = 0.9$.

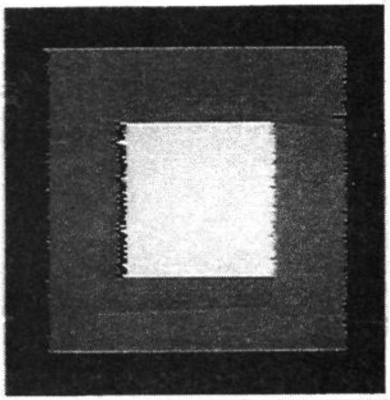

Fig 2: Maximum likelihood minimum discontinuity disparity map for rds.

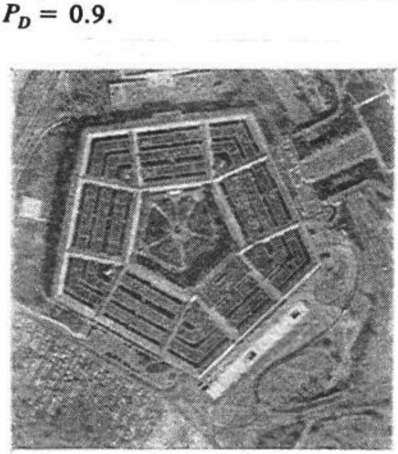

Fig 3: The Pentagon

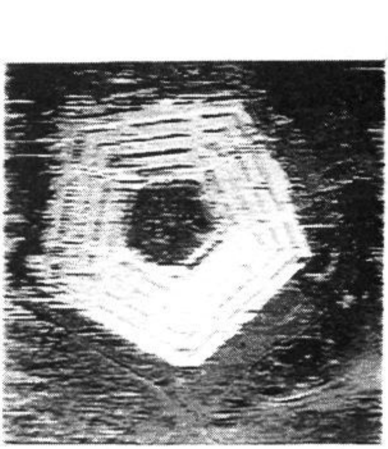

Fig 4: Maximum likelihood disparity map for the Pentagon for $P_D = 0.9$.

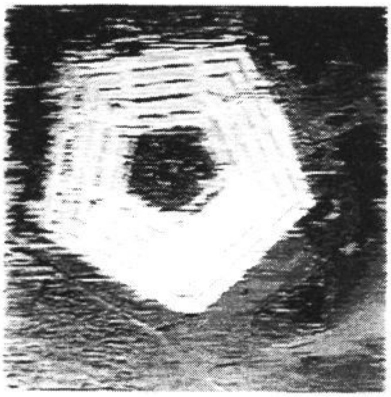

Fig 5: Maximum likelihood minimum discontinuity disparity map for the Pentagon for $P_D = 0.9$.

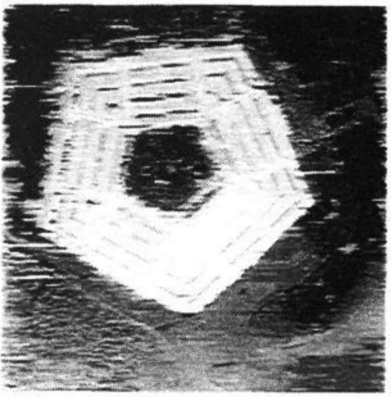

Fig 6: Maximum likelihood minimum discontinuity disparity map for the Pentagon for $P_D = 0.99$.

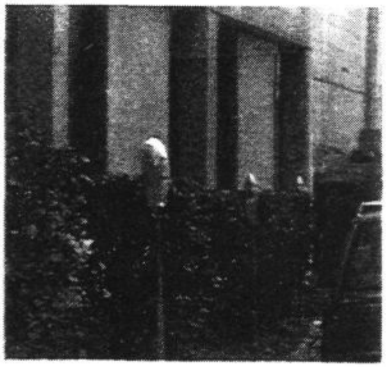

Fig 7: Left image of the "parked car" stereo pair.

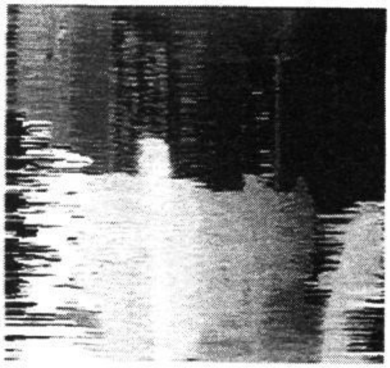

Fig 8: Maximum likelihood disparity map for the "Parked car".

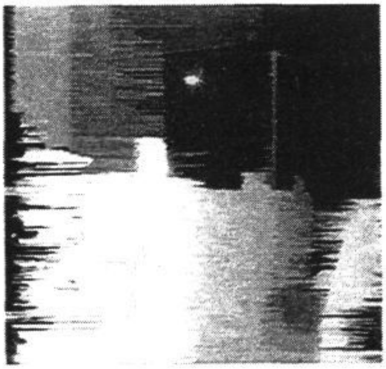

Fig 9: Maximum likelihood minimum discontinuity disparity map for the "parked Car".