A Computational Framework for Determining Stereo Correspondence from a Set of Linear Spatial Filters *

David G. Jones[1] *and Jitendra Malik*[2]

[1] McGill University, Dept. of Electrical Engineering, Montréal, PQ, Canada H3A 2A7
[2] University of California, Berkeley, Computer Science Division, Berkeley, CA USA 94720

Abstract. We present a computational framework for stereopsis based on the outputs of linear spatial filters tuned to a range of orientations and scales. This approach goes beyond edge-based and area-based approaches by using a richer image description and incorporating several stereo cues that have previously been neglected in the computer vision literature.

A technique based on using the pseudo-inverse is presented for characterizing the information present in a vector of filter responses. We show how in our framework viewing geometry can be recovered to determine the locations of epipolar lines. An assumption that visible surfaces in the scene are piecewise smooth leads to differential treatment of image regions corresponding to binocularly visible surfaces, surface boundaries, and occluded regions that are only monocularly visible. The constraints imposed by viewing geometry and piecewise smoothness are incorporated into an iterative algorithm that gives good results on random-dot stereograms, artificially generated scenes, and natural grey-level images.

1 Introduction

Binocular stereopsis is based on the cue of *disparity* — two eyes (or cameras) receive slightly different views of the three-dimensional world. This disparity cue, which includes differences in position, both horizontal and vertical, as well as differences in orientation or spacing of corresponding features in the two images, can be used to extract the three-dimensional structure in the scene. This depends, however, upon first obtaining a solution to the correspondence problem. The principal constraints that make this feasible are:

1. Similarity of corresponding features in the two views.
2. Viewing geometry which constrains corresponding features to lie on epipolar lines.
3. Piecewise continuity of surfaces in the scene because of which nearby points in the scene have nearby values of disparity. The disparity gradient constraint (Burt and Julesz, 1980; Pollard et al., 1985) and the ordering constraint (Baker and Binford, 1982) are closely related.

Different approaches to the correspondence problem exploit these constraints in different ways. The two best studied approaches are area correlation (Hannah, 1974; Gennery, 1977; Moravec, 1977; Barnard and Thompson, 1980) and edge matching (Marr and Poggio, 1979; Grimson, 1981; Baker and Binford, 1982; Pollard et al., 1985; Medioni and Nevatia, 1985; Ayache and Faverjon, 1987).

* This work has been supported by a grant to DJ from the Natural Sciences and Engineering Research Council of Canada (OGP0105912) and by a National Science Foundation PYI award (IRI-8957274) to JM.

The difficulties with approaches based on area correlation are well known. Because of the difference in viewpoints, the effects of shading can give rise to differences in brightness for non-lambertian surfaces. A more serious difficulty arises from the effects of differing amounts of foreshortening in the two views whenever a surface is not strictly fronto-parallel. Still another difficulty arises at surface boundaries, where a depth discontinuity may run through the region of the image being used for correlation. It is not even guaranteed in this case that the computed disparity will lie within the range of disparities present within the region.

In typical edge-based stereo algorithms, edges are deemed compatible if they are near enough in orientation and have the same sign of contrast across the edge. To cope with the enormous number of false matches, a coarse-to-fine strategy may be adopted (e.g., Marr and Poggio, 1979; Grimson, 1981). In some instances, additional limits can be imposed, such as a limit on the rate at which disparity is allowed to change across the image (Mayhew, 1983; Pollard et al., 1985). Although not always true, assuming that corresponding edges must obey a left-to-right ordering in both images can also be used to restrict the number of possible matches and lends itself to efficient dynamic programming methods (Baker and Binford, 1982). With any edge-based approach, however, the resulting depth information is sparse, available only at edge locations. Thus a further step is needed to interpolate depth across surfaces in the scene.

A third approach is based on the idea of first convolving the left and right images with a bank of linear filters tuned to a number of different orientations and scales (e.g., Kass, 1983). The responses of these filters at a given point constitute a vector that characterizes the local structure of the image patch. The correspondence problem can be solved by seeking points in the other view where this vector is maximally similar.

Our contribution in this paper is to develop this filter-based framework. We present techniques that exploit the constraints arising from viewing geometry and the assumption that the scene is composed of piecewise smooth surfaces. A general viewing geometry is assumed, with the optical axes converged at a fixation point, instead of the simpler case of parallel optical axes frequently assumed in machine vision. Exploiting piecewise smoothness raises a number of issues — the correct treatment of depth discontinuities, and associated occlusions, where unpaired points lie in regions seen only in one view. We develop an iterative framework (Fig. 1) which exploits all these constraints to obtain a dense disparity map. Our algorithm maintains a current best estimate of the viewing parameters (to constrain vertical disparity to be consistent with epipolar geometry), a visibility map (to record whether a point is binocularly visible or occluded), and a scale map (to record the largest scale of filter not straddling a depth discontinuity).

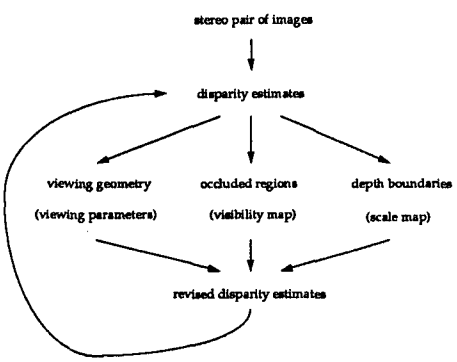

Fig. 1. Iteratively refining estimates of stereo disparity.

This paper is organized as follows. Section 2 gives an introduction to the use of filtering as a first stage of visual processing. A technique based on using the pseudo-inverse is presented for characterizing the information present in a vector of filter responses. Section 3 demonstrates the performance of a simple-minded matching strategy based on just comparing filter responses. This helps to motivate the need for exploiting the additional constraints imposed by the viewing geometry and piecewise smoothness. These constraints are developed further in Section 4. In section 5 the complete algorithm is presented. Section 6 concludes with experimental results.

2 Local Analysis of Image Patches by Filtering

In order to solve the correspondence problem, stereo algorithms attempt to match *features* in one image with corresponding features in the other. Central to the design of these algorithms are two choices: *What* are the image features to be matched? *How* are these features compared to determine corresponding pairs.

It is important to recall that stereo is just one of many aspects of early visual processing: stereo, motion, color, form, texture, etc. It would be impractical for each of these to have its own specialized representation different from the others. The choice of a "feature" to be used as the basis for stereopsis must thus be be constrained as a choice of the input representation for many early visual processing tasks, not just stereo. For the human visual system, a simple feature such as a "pixel" is not even available in the visual signals carried out of the eye. Already the pattern of light projected on the retina has been sampled and spatially filtered. At the level of visual inputs to the cortex, visual receptive fields are well approximated as linear spatial filters, with impulse response functions that are the Laplacian of a two-dimensional Gaussian, or simply a difference of Gaussians. Very early in cortical visual processing, receptive fields become oriented and are well approximated by linear spatial filters, with impulse response functions that are similar to partial derivatives of a Gaussian (Young, 1985).

Since "edges" are derived from spatial filter outputs, the detection and localization of edges may be regarded as an unnecessary step in solving the correspondence problem. A representation based on edges actually discards information useful in finding unambiguous matches between image features in a stereo pair. An alternative approach, explored here, is to treat the the spatial filter responses at each image location, collectively called the *filter response vector*, as the feature to be used for computing stereo correspondence.

Although this approach is loosely inspired by the current understanding of processing in the early stages of the primate visual system (for a recent survey, DeValois and DeValois, 1988), the use of spatial filters may also be viewed analytically. The filter response vector characterizes a local image region by a set of values at a point. This is similar to characterizing an analytic function by its derivatives at a point. From such a representation, one can use a Taylor series approximation to determine the values of the function at neighboring points. Because of the commutativity of differentiation and convolution, the spatial filters used are in fact computing "blurred derivatives" at each point. The advantages of such a representation have been described in some detail (Koenderink and van Doorn, 1987; Koenderink, 1988). Such a representation provides an efficient basis for various aspects of early visual processing, making available at each location of the computational lattice, information about a whole neighborhood around the point.

The primary goal in using a large number of spatial filters, at various orientations, phases, and scales is to obtain rich and highly specific image features suitable for stereo matching, with little chance of encountering false matches. At this point, one might be

tempted to formulate more precise, mathematical criteria and to attempt to determine an *optimal* set of filters. The alternative viewpoint taken here is that a variety of filter sets would each be adequate and any good stereo algorithm should not depend critically upon the precise form of the spatial filters chosen.

2.1 The Filter Set

The implementation and testing of these ideas requires some particular set of filters to be chosen, though at various times, alternative filters to those described below have been used, always giving more or less similar results. The set of filters used consisted of rotated copies of filters with impulse responses $F(x,y) = G_n(x) \times G_0(y)$, where $n = 1, 2, 3$ and G_n is the n^{th} derivative of a Gaussian. The scale, σ, was chosen to be the same in both the x and y directions. Filters at seven scales were used, with the area of the filters increasing by a factor of two at each scale. In terms of pixels, the filters are $w \times w$, with $w \in \{3, 5, 7, 10, 14, 20, 28\}$, and $w = \lceil 8\sigma \rceil$. The filters at the largest scale are shown in Fig. 2. Smaller versions of the same filters are used at finer scales. Nine filters at seven scales would give 63 filters, except at the finest scale the higher derivatives are useless because of quantization errors, and so were discarded.

2.2 Singular Value Decomposition

Regardless of why a particular set of filters may be chosen, it is useful to know that there is an automatic procedure that can be used to evaluate the degree to which the chosen filters are independent. Any filter that can be expressed as the weighted sum of others in the set is redundant. Even filters for which this is not strictly true, but *almost* true may be a poor choice, especially where this may lead to numerical instability in some computations involving filter responses. The singular value decomposition provides just this information.

Any $m \times n$ matrix A, may be expressed as the product of an $m \times m$ matrix U, an $m \times n$ diagonal matrix Σ, and an $n \times n$ matrix V^T, where the columns of U and V are orthonormal, and the entries in Σ are positive or zero. This decomposition is known as the *singular value decomposition*. The diagonal entries of the matrix Σ are called singular values and satisfy $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq 0$. More details may be found in a standard linear algebra or numerical analysis text (e.g., Golub and Van Loan, 1983).

A spatial filter with finite impulse response may be represented as an $n \times 1$ column vector, F_i, by writing out its entries row by row. Here n is the number of pixels in the support of the filter. If an image patch (of the same size and shape as the support of the filter) is also represented as an $n \times 1$ column vector, then the result of convolving the image patch by the filter is simply the inner product of these two vectors. Taken together, a set of spatial filters forms a matrix F. This is a convenient representation of the linear transformation that maps image patches to a vector of filter responses. For an image patch represented as a vector I, the filter response vector is simply $v = F^T I$. Applying the singular value decomposition yields $F^T = U \Sigma V^T$

The number of non-zero entries in Σ is the rank, r, or the dimension of the vector space spanned by the filters. The first r columns of V form an orthonormal basis set for this vector space, ranked in order of the visual patterns to which this particular set of filters is most sensitive. The corresponding singular values indicate how sensitive. The remaining columns form an orthonormal basis for the null space of F — those spatial patterns to which F is entirely insensitive. The matrix U may be thought of as an orthonormal basis set for the space of possible filter responses vectors, or merely as a

change of basis matrix. As an example of this decomposition, the orthonormal basis for the set of filters in Fig. 2A is shown in Fig. 2B.

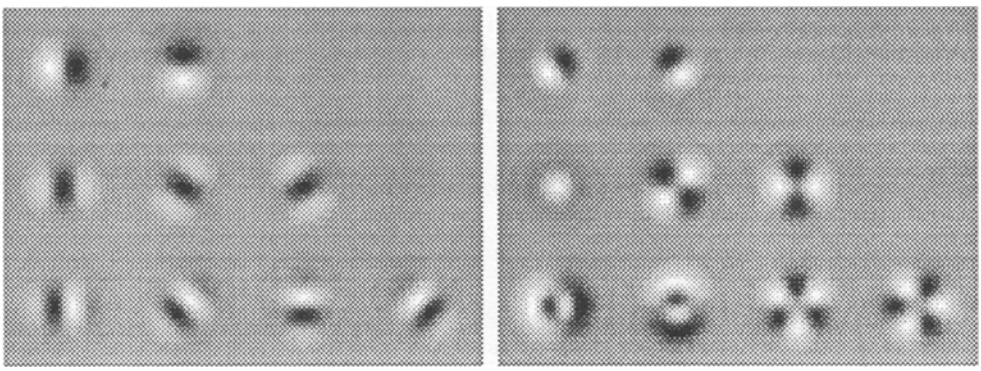

Fig. 2. *A.* Linear spatial filter set. *B.* Orthonormal basis set for vector space spanned by filters in *A*.

One telltale sign of a poorly chosen set of filters is the presence of singular values that are zero, or very close to zero. Consider, for example, a filter set consisting of the first derivative of a Gaussian at four different orientations, θ.

$$G^{\theta}_{1,0}(x,y) = G_1(u) \times G_0(v) \quad ; \quad u = x\cos\theta - y\sin\theta, \quad v = x\sin\theta + y\cos\theta$$

The vector space spanned by these four filters is only two dimensional. Only two filters are needed, since the other two may be expressed as the weighted sum of these, and thus carry no additional information. If one did not already know this analytically, this procedure quickly makes it apparent. Such filters for which responses at a small number of orientations allow the easy computation of filter responses for other orientations have been termed *steerable filters* (Koenderink, 1988; Freeman and Adelson, 1991; Perona, 1991). For Gaussian derivatives in particular, it turns out that $n+1$ different orientations are required for the n^{th} Gaussian derivative.

As a further example, the reader who notes the absence of unoriented filters in Fig. 2A and is tempted to enrich the filter set by adding a $\nabla^2 G$, Laplacian of Gaussian filter, should think twice. This filter is already contained in the filter set in the sense that it may be expressed as the weighted sum of the oriented filters $G^{\theta}_{2,0}(x,y)$. Similar filters, such as a difference of Gaussians, may not be entirely redundant, but they result in singular values close to zero, indicating that they add little to the filter set.

At the coarsest scales, filter responses vary quite smoothly as one moves across an image. For this reason, the filter response at one position in the image can quite accurately be computed from filter responses at neighboring locations. This means it is not strictly necessary to have an equal number of filters at the coarser scales, and any practical implementation of this approach would take advantage of this by using progressively lower resolution sampling for the larger filter scales. Regardless of such an implementation decision, it may be assumed that the output of every filter in the set is available at every location in the image, whether it is in fact available directly or may be easily computed from the outputs of a lower resolution set of filters.

2.3 Image Encoding and Reconstruction

What information is actually carried by the filter response vector at any given position in an image? This important question is surprisingly easy to answer. The singular value decomposition described earlier provides all that is necessary for the best least-squares reconstruction of an image patch from its filter response vector. Since $v = F^T I$, and $F^T = U \Sigma V^T$, the reconstructed image patch can be computed using the *generalized inverse* (or the Moore-Penrose pseudo-inverse) of the matrix F^T.

$$I' = V \, 1/\Sigma \, U^T \, v$$

The matrix $1/\Sigma$ is a diagonal matrix obtained from Σ by replacing each non-zero diagonal entry σ_i by its reciprocal, $1/\sigma_i$.

An example of such a reconstruction is given in Fig. 3. The finest detail is preserved in the center of the patch where the smallest filters are used. The reconstruction is progressively less accurate as one moves away from from the center. Because there are fewer filters than pixels in the image patch to be reconstructed, the reconstruction is necessarily incomplete. The high quality of the the reconstructed image, however, confirms the fact that most of the visually salient features have been preserved. The reduction in the number of values needed to represent an image patch means this is an efficient encoding — not just for stereo, but for other aspects of early visual processing in general. Since this same encoding is used throughout the image, this notion of efficiency should be used with caution. In terms of merely representing the input images, storing a number of filter responses for each position in the image is clearly less efficient than simply storing the individual pixels. In terms of carrying out computations on the image, however, there is a considerable savings for even simple operations such as comparing image patches. Encoded simply as pixels, comparing 30 × 30 image regions requires 900 comparisons. Encoded as 60 filter responses, the same computation requires one-fifteenth as much effort.

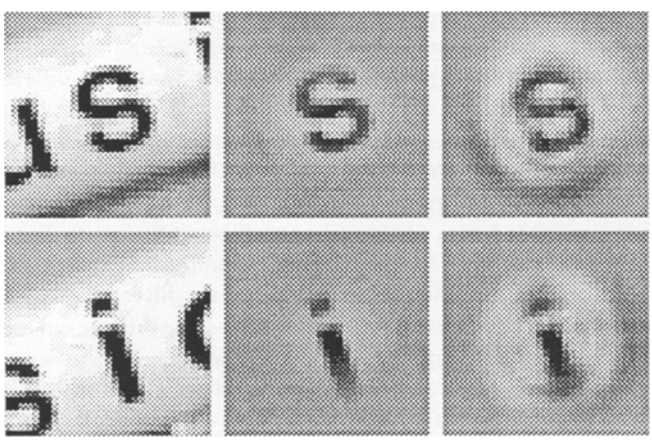

Fig. 3. Image reconstruction. Two example image patches (*left*), were reconstructed (*right*) from spatial filter responses at their center. Original image patches masked by a Gaussian (*middle*) are shown for comparison.

3 Using Filter Outputs for Matching

How should filter response vectors be compared? Although corresponding filter response vectors in the two views should be very similar, differences in foreshortening and shading mean that they will rarely be identical. A variety of measures can be used to compare two vectors, including the angle between them, or some norm of their vector difference. These and similar measures are zero when the filter response vectors are identical and otherwise their magnitude is proportional to some aspect of the difference between potentially corresponding image patches. It turns out that any number of such measures do indistinguishably well at identifying corresponding points in a pair of stereo images, except at depth discontinuities. Near depth discontinuities, the larger spatial filters lie across an image patch containing the projection of more than one surface. Because these surfaces lie at different depths and thus have different horizontal disparities, the filter responses can differ considerably in the two views, even when they are centered on points that correspond. While the correct treatment of this situation requires the notion of an adaptive scale map (developed in the next section), it is helpful to use a measure such as the L_1 norm, the sum of absolute differences of corresponding filter responses, which is less sensitive to the effect of such outliers than the L_2 norm.

$$e_m = \sum_k |F_k * I_r(i,j) - F_k * I_l(i + h_r, j + v_r)|$$

This matching error e_m is computed for a set of candidate choices of (h_r, v_r) in a window determined by a priori estimates of the range of horizontal and vertical disparities. The (h_r, v_r) value that minimizes this expression is taken as the best initial estimate of positional disparity at pixel (i,j) in the right view. This procedure is repeated for each pixel in both images, providing disparity maps for both the left and right views. Though these initial disparity estimates can be quite accurate, they can be substantially improved using several techniques described in the next section.

An implementation of this approach using the outputs of a number of spatial filters at a variety of orientations and scales as the basis for establishing correspondence has proven to give quite good results, for random-dot stereograms, as well as natural and artificial grey-level images. Some typical examples are presented here.

The recovered disparity map for a Julesz random-dot stereogram is presented in Fig. 4A. The central square standing out in depth is clearly detected. Disparity values at each image location are presented as grey for zero horizontal disparity, and brighter or darker shades for positive or negative disparities. Because these are offsets in terms of image coordinates, the disparity values for corresponding points in the left and right images should have equal magnitudes, but opposite signs. Whenever the support of the filter set lies almost entirely on a single surface, the disparity estimates are correct. Even close to depth discontinuities, the recovered disparity is quite accurate, despite the responses from some of the larger filters being contaminated by lying across surfaces at different depths.

In each view, there is a narrow region of the background just to one side of the near central square that is visible only in one eye. In this region, there is no corresponding point in the other view and the recovered disparity estimates appear as noise. Methods for coping with these initial difficulties are discussed in later sections. In the lower panels of the same figure, the measure of dissimilarity, e_m, between corresponding filter response vectors is shown, with darker shades indicating larger differences. Larger differences are clearly associated with depth discontinuities.

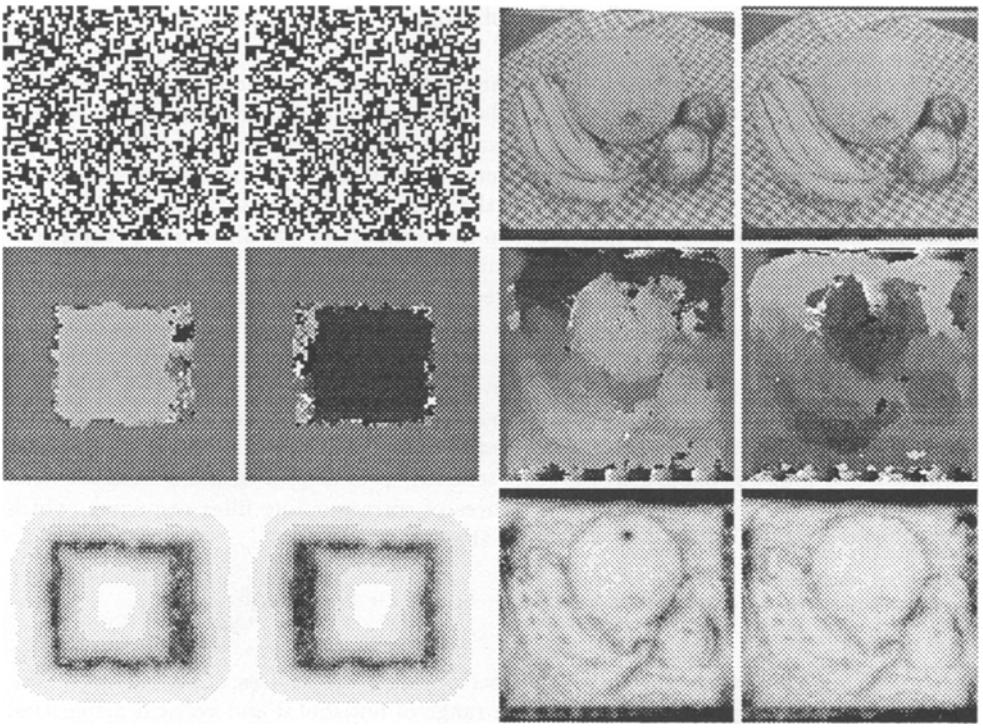

Fig. 4. Initial disparity estimates: random-dot stereogram and fruit. For the stereo pairs shown (*top*), the recovered disparity map (*middle*) and dissimilarity or error map (*bottom*) are shown. (fruit images courtesy Prof. N. Ahuja, Univ. Illinois)

When approached as a problem of determining which black dot in one view corresponds with which black dot in the other, the correspondence problem seems quite difficult. In fact, Julesz random-dot stereograms are among the richest stimuli — containing information at all orientations and scales. When the present approach based on spatial filters is used, the filter response vector at each point proves to be quite distinctive, making stereo-matching quite straightforward and unambiguous.

As an example of a natural grey-level image, a stereo pair of fruit lying on a table cloth is shown in Fig. 4B. The recovered disparity values clearly match the shapes of the familiar fruit quite well. Once again, some inaccuracies are present right at object boundaries. The measure of dissimilarity, or error shown at the bottom of the figure provides a blurry outline of the fruit in the scene. A mark on the film, present in one view and not the other (on the canteloupe) is also clearly identified in this error image.

As a final example, a ray-traced image of various geometric shapes in a three-sided room is depicted in Fig. 5. For this stereo pair, the optical axes are not parallel, but converged to fall on a focal point in the scene. This introduces vertical disparities between corresponding points. Estimated values for both the horizontal and vertical disparities are shown. Within surfaces, recovered disparities values are quite accurate and there are some inaccuracies right at object boundaries. Just to the right of the polyhedron in this scene is a region of the background visible only in one view. The recovered disparity values are nonsense, since even though there is no correct disparity, this method will always choose one candidate as the "best". Another region in this scene where there

are some significant errors is along the room's steeply slanted left wall. In this case, the large differences in foreshortening between the two views poses a problem, since the filter responses at corresponding points on this wall will be considerably different. A method for handling slanted surfaces such as this has been discussed in detail elsewhere (Jones, 1991; Jones and Malik, 1992).

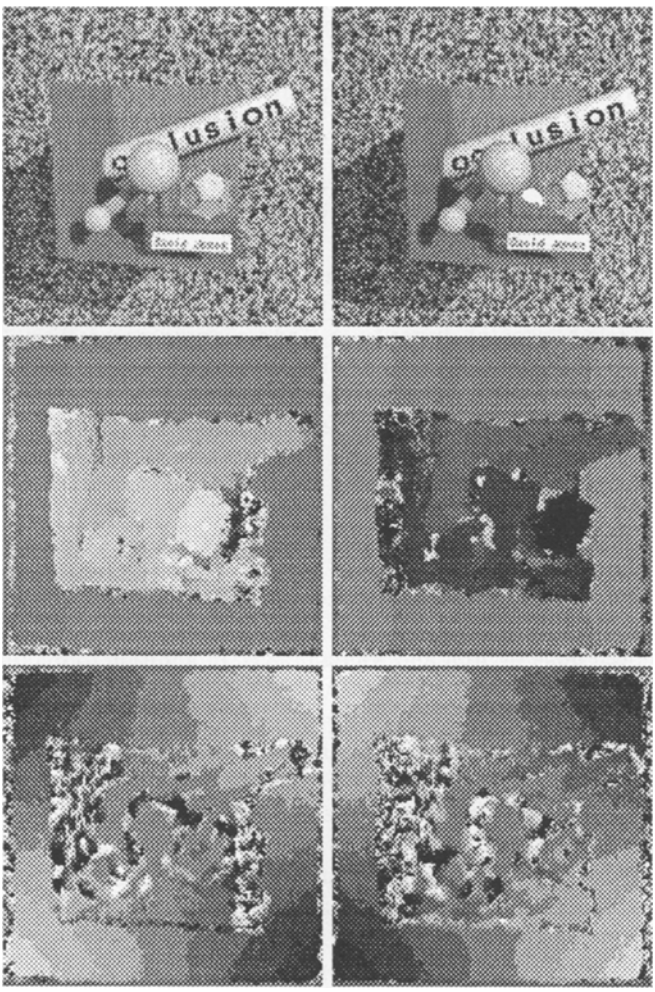

Fig. 5. Initial disparity estimates: a simple raytraced room. For the stereo pair (*top*), the recovered estimates of the horizontal (*middle*) and vertical (*bottom*) components of positional disparity are shown.

4 Additional constraints for solving correspondence

4.1 Epipolar Geometry

By virtue of the basic geometry involved in a pair of eyes (or cameras) viewing a three-dimensional scene, corresponding points must always lie along epipolar lines in the images. These lines correspond to the intersections of an epipolar plane (the plane through

a point in the scene and the nodal points of the two cameras) with the left and right image planes. Exploiting this epipolar constraint reduces an initially two-dimensional search to a one-dimensional one. Obviously determination of the epipolar lines requires a knowledge of the viewing geometry.

The core ideas behind the algorithms to determine viewing geometry date back to work in the photogrammetry community in the beginning of this century (for some historical references, Faugeras and Maybank, 1990) and have been rediscovered and developed in the work on structure from motion in the computational vision community. Given a sufficient number of corresponding pairs of points in two frames (at least five), one can recover the rigid body transformation that relates the two camera positions except for some degenerate configurations. In the context of stereopsis, Mayhew (1982) and Gillam and Lawergren (1983) were the first to point out that the viewing geometry could be recovered purely from information present in the two images obtained from binocular viewing.

Details of our algorithm for estimating viewing parameters may be found in (Jones and Malik, 1991). We derive an expression for vertical disparity, v_r, in terms of image coordinates, (i_r, j_r), horizontal disparity, h_r, and viewing parameters. This condition must hold at all positions in the image, allowing a heavily over-constrained determination of certain viewing parameters. With the viewing geometry known, the image coordinates and horizontal disparity *determine* the vertical disparity, thus reducing an initially two-dimensional search for corresponding points to a one-dimensional search.

4.2 Piecewise smoothness

Since the scene is assumed to consist of piecewise smooth surfaces, the disparity map is piecewise smooth. Exploiting this constraint requires some subtlety. Some previous work in this area has been done by Hoff and Ahuja (1989). In addition to making sure that we do not smooth away the disparity discontinuities associated with surface boundaries in the scene, we must also deal correctly with regions which are only monocularly visible.

Whenever there is a surface depth discontinuity which is not purely horizontal, distant surfaces are occluded to different extents in the two eyes, leading to the existence of unpaired image points which are seen in one eye only. The realization of this goes back to Leonardo Da Vinci (translation in, Kemp, 1989). This situation is depicted in Fig. 6.

Recent psychophysical work has convincingly established that the human visual system can exploit this cue for depth in a manner consistent with the geometry of the situation (Nakayama and Shimojo, 1990).

Any computational scheme which blindly assigns a disparity value to each pixel is bound to come up with nonsense estimates in these regions. Examples of this can be found by inspecting the occluded regions in Fig. 5. At the very minimum, the matching algorithm should permit the labeling of some features as 'unmatched'. This is possible in some dynamic programming algorithms for stereo matching along epipolar lines (e.g., Arnold and Binford, 1980) where vertical and horizontal segments in the path through the transition matrix correspond to skipping features in either the left or right view.

In an iterative framework, a natural strategy is to try and identify at each stage the regions which are only monocularly visible. The hope is that while initially this classification will not be perfect (some pixels which are binocularly visible will be mislabeled as monocularly visible and vice versa), the combined operation of the different stereopsis constraints would lead to progressively better classification in subsequent iterations. Our empirical results bear this out.

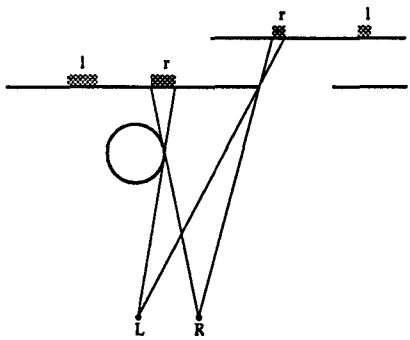

Fig. 6. Occlusion. In this view from above, it is clear that at depth discontinuities there are often regions visible to one eye, but not the other. To the right of each near surface is a region r that is visible only to the right eye, R. Similarly, to the left of a near surface is monocular region, l, visible only to the left eye, L.

The problem of detecting and localizing occluded regions in a pair of stereo images is made much easier when one recalls that there are indeed a *pair* of images. The occluded regions in one image include exactly those points for which there is no corresponding point in the other image. This suggests that the best cue for finding occluded regions in one image lies in the disparity estimates for the *other* image!

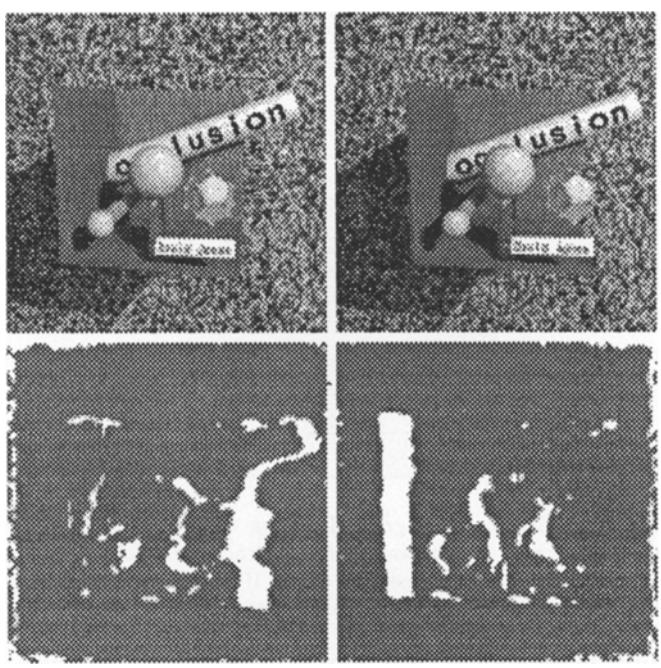

Fig. 7. Visibility map. The white areas in the lower panels mark the regions determined to be visible only from one of the two viewpoints.

Define a binocular *visibility map*, $B(i,j)$, for one view as being 1 at each image position that is visible in the other view, and 0 otherwise (i.e., an occluded region). The

horizontal and vertical disparity values for each point in, say, the left image are signed offsets that give the coordinates of the corresponding point in the right image. If the visibility map for the right image is initially all zero, it can be filled in systematically as follows. For each position in the left image, set the corresponding position in the right visibility map to 1. Those positions that remain zero had no corresponding point in the other view and are quite likely occluded. An example of a visibility map computed in this manner is shown in Fig. 7.

Having established a means for finding regions visible only from a one viewpoint, what has been achieved? If the disparity values are accurate, then the visibility map, besides simply identifying binocularly visible points, also explicitly delimits occluding contours. After the final iteration, occluded regions can be assigned the same disparity as the more distant neighboring visible surface.

4.3 Depth Discontinuities and Adaptive Scale Selection

The output of a set of spatial filters at a range of orientations and scales provides a rich description of an image patch. For corresponding image patches in a stereo pair of images, it is expected that these filter outputs should be quite similar. This expectation is reasonable when all of the spatial filters are applied to image patches which are the projections of single surfaces. When larger spatial filters straddle depth discontinuities, possibly including occluded regions, the response of filters centered on corresponding image points may differ quite significantly. This situation is depicted in Fig. 8. Whenever a substantial area of a filter is applied to a region of significant depth variation, this difficulty occurs (e.g., in Fig. 5).

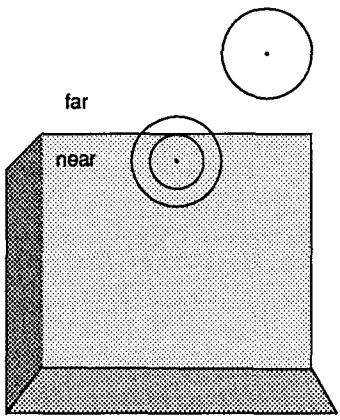

Fig. 8. Scale selection. Schematic diagram depicting a three-sided room similar to the one in Fig. 5. When attempting to determining correspondence for a point on a near surface, larger filters that cross depth boundaries can result in errors. If depth discontinuities could be detected, such large scale filters could be selectively ignored in these situations.

From an initial disparity map, it is possible to estimate where such inappropriately large scale filters are being used by applying the following procedure. At each position in the image, the median disparity is determined over a neighborhood equal to the support of the largest spatial filter used for stereo matching. Over this same neighborhood, the difference between each disparity estimate and this median disparity is determined. These differences are weighted by a Gaussian at the same scale as the filter, since the center of

the image patch has a greater effect on the filter response. The sum of these weighted disparity differences provides a measure of the amount of depth variation across the image patch affecting the response of this spatial filter. When this sum exceeds an appropriately chosen threshold, it may be concluded that the filter is too large for its response to be useful in computing correspondence. Otherwise, continuing to make use of the outputs of large spatial filters provides stability in the presence of noise.

To record the results of applying the previous procedure, the notion of a *scale map* is introduced (Fig. 9). At each position in an image, the scale map, $S(i, j)$, records the scale of the largest filter to be used in computing stereo correspondence. For the computation of initial disparity estimates, all the scales of spatial filters are used. From initial disparity estimates, the scale map is modified using the above criterion. At each position, if it is determined that an inappropriately large scale filter was used, then the scale value at that position is decremented. Otherwise, the test is redone at the next larger scale, if there is one, to see if the scale can be incremented. It is important that this process of adjusting the scale map is done in small steps, with the disparity values being recalculated between each step. This prevents an initially noisy disparity map, which seems to have a great deal of depth variation, from causing the largest scale filters to be incorrectly ignored.

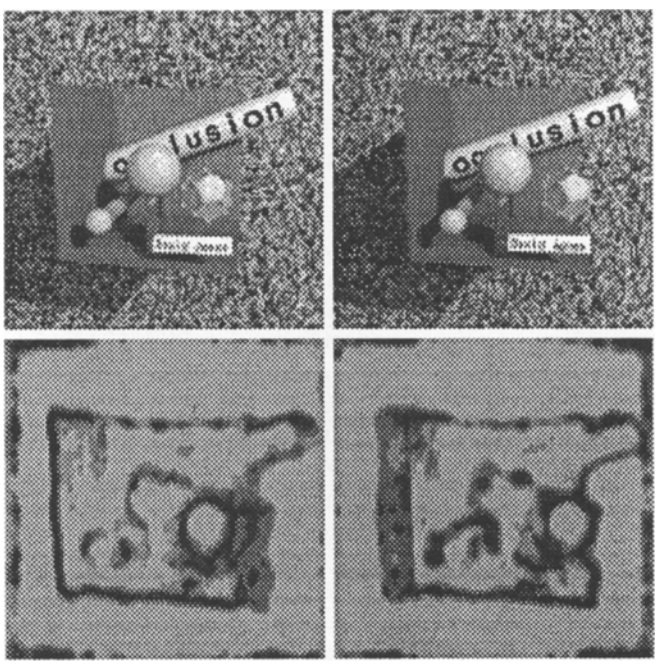

Fig. 9. Scale map. The darker areas in the lower panels mark the regions where larger scale filters are being discarded because they lie across depth discontinuities.

5 The Complete Algorithm

Once initial estimates of horizontal and vertical disparity have been made, additional information becomes available which can be used to improve the quality of the disparity estimates. This additional information includes estimates of the viewing parameters, the

location of occluded regions, and the appropriate scale of filters to be used for matching. Our algorithm can be summarized as follows:

1. For each pixel P with coordinates (i, j) in the *left* image, and for each candidate disparity value $\hat{h}, \hat{v}$ in the allowable disparity range compute the error measure $e_{ij}(\hat{h}, \hat{v})$.
2. Declare $h(i, j)$ and $v(i, j)$ to be the values of $\hat{h}, \hat{v}$ that minimize e_{ij}.
3. Use the refined values of $h(i, j)$ and $v(i, j)$ to compute the new visibility map $B(i, j)$ and scale map $S(i, j)$.
4. Perform steps 1–3 for disparity, visibility, and scale maps but this time with respect to the *right* image.
5. Goto step 1 or else stop at convergence.

The error function $e(\hat{h}, \hat{v})$ is the sum of the following terms

$$e(\hat{h}, \hat{v}) = \lambda_m e_m(\hat{h}, \hat{v}) + \lambda_v e_v(\hat{h}, \hat{v}) + \lambda_c e_c(\hat{h}, \hat{v}) + \lambda_s e_s(\hat{h}, \hat{v})$$

Each term enforces one of the constraints discussed: similarity, viewing geometry, consistency, and smoothness. The λ parameters control the weight of each of these constraints, and their specific values are not particularly critical. The terms are:

- $e_m(\hat{h}, \hat{v})$ is the matching error due to dissimilarity of putative corresponding points. It is 0 if $B(i, j) = 0$ (i.e., the point is occluded in the other view), otherwise it is $\sum_k |F_k * I_r(i, j) - F_k * I_l(i + h_r, j + v_r)|$ where k ranges from the smallest scale to the scale specified by $S(i, j)$.
- $e_v(\hat{h}, \hat{v})$ is the vertical disparity error $|\hat{v} - v^*|$ where v^* is the vertical disparity consistent with the recovered viewing parameters. This term enforces the epipolar geometry constraint.
- $e_c(\hat{h}, \hat{v})$ is the consistency error between the disparity maps for the left and right images. Recall that in our algorithm the left and right disparity maps are computed independently. This term provides the coupling — positional disparity values for corresponding points should have equal magnitudes, but opposite signs. If h', v' is the disparity assigned to the corresponding point $P' = (i + \hat{h}, j + \hat{v})$ in the other image, then $h' = -\hat{h}$ and $v' = -\hat{v}$ at binocularly visible points. If only one of P and P' is labelled as monocularly visible, then this is consistent only if the horizontal disparities place this point *further* than the binocularly visible point. In this case, $e_c = 0$, otherwise, $e_c = |\hat{h} + h'| + |\hat{v} + v'|$.
- $e_s(\hat{h}, \hat{v}) = |\hat{h} - \bar{h}| + |\hat{v} - \bar{v}|$ is the smoothness error used to penalize candidate disparity values that deviate significantly from $\bar{h}, \bar{v}$, the 'average' values of horizontal and vertical disparity in the neighborhood of P. These are computed either by a local median filter, within binocularly visible regions, or by a local smoothing operation within monocularly visible regions. These operations preserve boundaries of binocularly visible surfaces while providing stable depth estimates near occluded regions.

The computational complexity of this algorithm has two significant terms. The first is the cost of the initial linear spatial filtering at multiple scales and orientations. Implementations can be made quite efficient by using separable kernels and pyramid strategies. The second term corresponds to the cost of computing the disparity map. This cost is proportional to the number of iterations (typically 10 or so in our examples). The cost in each iteration is dominated by the search for the pixel in the other view with minimum e. This is $O(n^2 w_h w_v)$ for images of size $n \times n$ and horizontal and vertical disparity ranges, w_h and w_v. After the first iteration, when the viewing parameters have been estimated, the approximate vertical disparity is known at each pixel. This enables w_v to be restricted to be 3 pixels which is adequate to handle quantization errors of ± 1 pixel.

6 Experimental Results

The algorithm described in the previous section has been implemented and tested on a variety of natural and artificial images. In practice, this process converges (i.e., stops producing significant changes) in under ten iterations. Disparity maps obtained using this algorithm are shown in Fig. 10. The reader may wish to compare these with Figures 4 and 5 which show the disparity map after a single iteration when the correspondence is based solely on the similarity of the filter responses. The additional constraints of epipolar geometry and piecewise smoothness have clearly helped, particularly in the neighborhood of depth discontinuities. Also note that the visibility map for the random dot stereogram as well as the room image (bottom of Fig. 7) are as expected. From these representations, the detection and localization of depth discontinuities is straightforward.

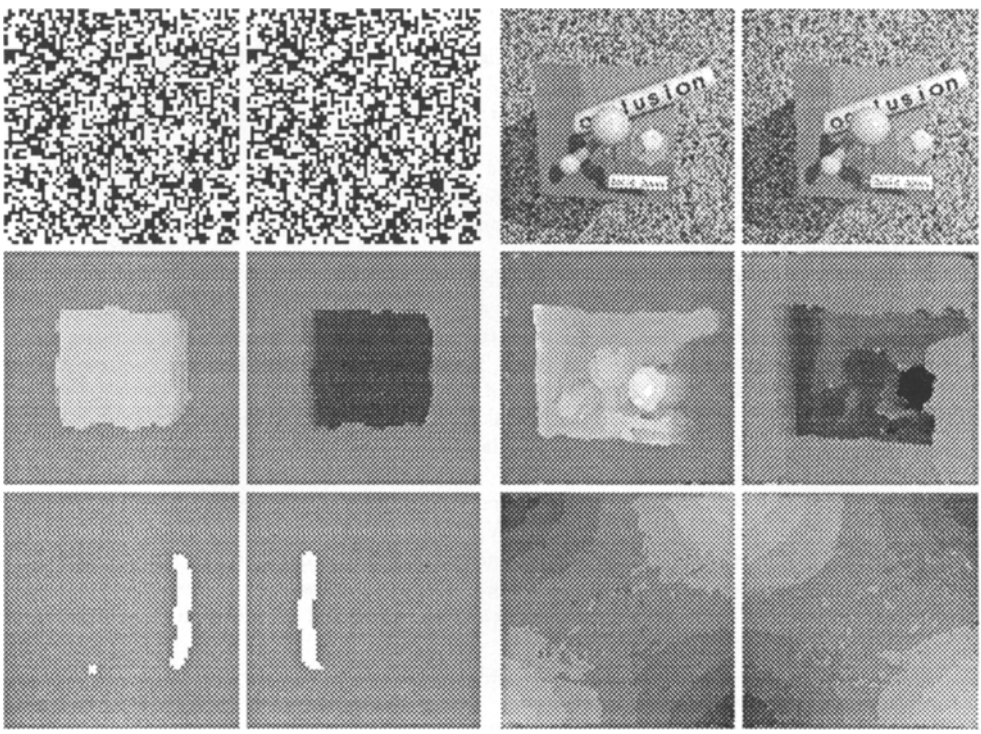

Fig. 10. Refined disparity estimates. For the stereo pairs (*top*), the recovered horizontal disparities are shown in the middle panel. For the random dot stereogram, the lower panel shows the visibility map. For the room image, the bottom panel shows the recovered vertical disparity.

We have demonstrated in this paper that convolution of the image with a bank of linear spatial filters at multiple scales and orientations provides an excellent substrate on which to base an algorithm for stereopsis, just as it has proved for texture and motion analysis. Starting out with a much richer description than edges was extremely useful for solving the correspondence problem. We have developed this framework further to enable the utilization of the other constraints of epipolar geometry and piecewise smoothness as well.

References

Arnold RD, Binford TO (1980) Geometric constraints on stereo vision. Proc SPIE 238:281–292

Ayache N, Faverjon B (1987) Efficient registration of stereo images by matching graph descriptions of edge segments. Int J Computer Vision 1(2):107–131

Baker HH, Binford TO (1981) Depth from edge- and intensity-based stereo. Proc 7th IJCAI 631–636

Barnard ST, Thompson WB (1980) Disparity analysis of images. IEEE Trans PAMI 2(4):333–340

Burt P, Julesz B (1980) A disparity gradient limit for binocular function. Science 208:651–657

DeValois R, DeValois K (1988) Spatial vision. Oxford Univ Press

Faugeras O, Maybank S (1990) Motion from point matches: multiplicity of solutions. Int J Computer Vision 4:225–246

Freeman WT, Adelson EH (1991) The design and use of steerable filters. IEEE Trans PAMI 13(9):891–906

Gennery DB (1977) A stereo vision system for autonomous vehicles. Proc 5th IJCAI 576–582

Gillam B, Lawergren B (1983) The induced effect, vertical disparity, and stereoscopic theory. Perception and Psychophysics 36:559–64

Golub GH, Van Loan CF (1983) Matrix computations. The Johns Hopkins Univ Press, Baltimore, MD

Grimson WEL (1981) From images to surfaces. M.I.T Press, Cambridge, Mass

Hannah MJ (1974) Computer matching of areas in images. Stanford AI Memo #239

Hoff W, Ahuja N (1989) Surfaces from stereo: integrating stereo matching, disparity estimation and contour detection. IEEE Trans PAMI 11(2):121–136

Jones, DG (1991) Computational models of binocular vision. PhD Thesis, Stanford Univ

Jones DG, Malik J (1991) A computational framework for determining stereo correspondence from a set of linear spatial filters. U.C. Berkeley Technical Report UCB–CSD 91–655

Jones DG, Malik J (1992) Determining three-dimensional shape from orientation and spatial frequency disparities. Proc ECCV, Genova

Kass M (1983) Computing visual correspondence. DARPA IU Workshop 54–60

Kemp M (Ed) (1989) Leonardo on painting. Yale Univ. Press: New Haven 65–66

Koenderink JJ, van Doorn AJ (1987) Representation of local geometry in the visual system. Biol Cybern 55:367–375

Koenderink JJ (1988) Operational significance of receptive field assemblies. Biol Cybern 58:163–171

Marr D, Poggio T (1979) A theory for human stereo vision. Proc Royal Society London B 204:301–328

Mayhew JEW (1982) The interpretation of stereo disparity information: the computation of surface orientation and depth. Perception 11:387–403

Mayhew JEW (1983) Stereopsis. in Physiological and Biological Processing of Images. Braddick OJ, Sleigh AC (Eds) Springer-Verlag, Berlin.

Medioni G, Nevatia R (1985) Segment-based stereo matching. CVGIP 31:2–18

Moravec HP (1977) Towards automatic visual obstacle avoidance. Proc 5th IJCAI

Nakayama K, Shimojo S (1990) DaVinci Stereopsis: Depth and subjective occluding contours from unpaired image points Vision Research 30(11):1811–1825

Perona P (1991) Deformable kernels for early vision. IEEE Proc CVPR 222–227

Pollard SB, Mayhew JEW, Frisby JP (1985) PMF: a stereo correspondence algorithm using a disparity gradient limit. Perception 14:449–470

Young R (1985) The Gaussian derivative theory of spatial vision: analysis of cortical cell receptive field line-weighting profiles. General Motors Research TR #GMR-4920