Detection of Binocular Disparities

K. Prazdny
Artificial Intelligence Laboratory, Schlumberger Palo Alto Research Center, 3340 Hillview Avenue, Palo Alto, CA 94304, USA

Abstract. A stereo correspondence algorithm designed to perform matching on figurally similar images (arising in normal human binocular vision) is described. It is based on the observation that the operational principles underlying biological stereo disparity detection seem to be extremely general and few in number instead of an extended set of specific “constraints”. We identify one general characteristic of objects in the three-dimensional world and use it to formulate a simple noniterative, parallel and local algorithm that successfully detects disparities generated by opaque as well as transparent surfaces.

Introduction

Binocular disparity is the difference between the positions of an object on the projection surfaces of the two eyes. It has been known since the invention of the stereoscope by Wheatstone (1838) that these disparities form the basis of binocular depth perception. Julesz’s (1960) invention of random-dot stereogram has demonstrated that perception of stereoscopic depth is subserved by a “primitive” mechanism that does not depend on the monocular recognition and identification of shapes, objects, or other high level cues. The information about three-dimensional surfaces in these stimuli is contained only in local correlations between elements in the two images because surfaces and shapes in such stereograms appear only after stereopsis has been achieved: each half of the stereo pair exhibits only (more or less isotropic) random texture. Such displays represent probably the ultimate form of camouflage.

Random-dot stereograms exemplify more than anything else the massive ambiguity problem the visual system may face: each picture element in one eye’s view can potentially correspond to many elements in the other eye’s view. In fact, the correspondence problem looks even more formidable because the human visual system can interpret random-dot stereograms portraying transparent surfaces (Fig. 1). Transparent surfaces are a disparity domain analogue of co-existing multiple organizations in other visual domains (see e.g. Prazdny, 1984).

As with many other capabilities, it is now widely recognized that stereopsis is not only a research area in psychology, physiology and psychophysics but also in information processing. In this report, the problem of stereopsis is approached primarily as a complex information processing task. Although our discussion does not directly bear on the question of how a physiological mechanism may detect binocular disparities, it addresses issues which may be, we believe, common to all such mechanisms, biological or artificial.

To detect the binocular disparity the visual system has to determine which location in one image corresponds to a given location in the other image. Considered as a computational problem, this involves the answer to two questions: what to match (i.e. what are the matching primitives) and how to match (i.e. how to discover the mapping from one image to the other). Correspondence between two images can be established by matching specific features such as blobs or edges, or by matching small regions by direct correlation of image intensities without identifying features. The basic problem with direct intensity correlation has to do with the fact that things can look significantly different from different points of view. Attempts to match directly the intensity values have had limited success and are not considered to be a biologically viable hypothesis. For example, Julesz (1971) has shown that a stereo pair consisting of images with different contrast (but the same contrast polarity!) can be easily fused. Surface markings and discontinuities are more invariant with respect to the change in viewing direction. It is thus not surprising that most computationally oriented theories (Sperling, 1970;
Fig. 1. A random dot stereogram portraying transparent surfaces. Stereopsis is easily obtained.

Dev, 1975; Marr and Poggio, 1976, 1979; Baker and Binford, 1981; Mayhew and Frisby, 1980) use edges or some other more primitive edge precursors as the main matching primitives. Recently, some success has been reported with a correlational technique based on a set of functions of intensity values (i.e. a vector) rather than directly on the (scalar) intensity values themselves (Kass, 1984).

The question of what matching primitives are used in human stereopsis is still largely open although it is known from neurophysiological observations that binocular cortical cells respond to oriented edges. In this report we assume that the descriptors have a punctiform nature, i.e. that they can be more or less accurately localized on the projection surface. One advantage of such descriptions is that they can be thought of as carrying only positional information. This property enables one to decouple the question of matching primitives from the problem of using them to detect binocular disparities and allows us to study the matching process by itself. The analysis and the algorithm developed below does not specify the type of matching primitives to be used.

An Analysis of the Problem

A computational approach to the problem begins with an analysis of the domain in which the stereoscopic mechanism operates: the real physical world (Gibson, 1950; Marr, 1982). Physical objects exhibit one important general property relevant to our purpose: they are "cohesive", i.e. surfaces of objects are smooth relative to the viewing distance (Marr and Poggio, 1976). From the cohesivity of matter follows directly the coherence principle: the world is not made of points chaotically varying in depth but of (not necessarily opaque) objects each occupying a well-defined 3D volume. The principle is different from the continuity constraints used implicitly (Sperling, 1970; Dev, 1975; Nelson, 1975) or explicitly (Marr and Poggio, 1976, 1979; Baker and Binford, 1981; Julesz, 1971; Mayhew, 1983) in most previous theories. In fact, the cohesivity of matter cannot be used to support the Marr and Poggio (1976, 1979) type of continuity rule for processing disparities arising from transparent surfaces. The continuity rule stipulates that nearby image points are projections of nearby 3D points. This implies smoothness (in the usual sense) of the resultant disparity field. Such continuity holds, however, only for opaque surfaces and non-boundary regions. The coherence principle is much more general. It recognizes that for transparent surfaces where proximal points on the projection surface may arise from widely separated three-dimensional objects, image proximity does not necessarily imply disparity continuity. While the disparity field may be locally discontinuous, it must (if it is generated by an actual three-dimensional scene obeying the coherence principle) be a superposition of locally smooth disparity fields corresponding to individual three-dimensional surfaces. These smooth variations usually are apparent only when larger image regions are taken into consideration. Locally, the field may be discontinuous due to disparities originating at different depth. In short, a discontinuous disparity field may be a superposition of a number of several interlaced continuous disparity fields each corresponding to a piecewise smooth surface. The coherence principle captures this possible state of affairs and includes continuous disparity variations associated with opaque surfaces as a special case.

The disparity detection problem can now be formulated as a decision problem. The visual system has to choose, from a set of possible disparities at a retinal position, the disparity which explains the best (in some sense) the distribution of possible correspondences generated by points (features) in a given neighborhood. In our current work, the set of possible disparities is further restricted to matches within an area of fixed radius (the area of local stereopsis) along the epipolar line. The coherence principle requires that neighbouring disparities of elements corresponding to the

---

1 The continuity constraint states that disparity varies smoothly almost everywhere and that only a small fraction of the area of an image is composed of boundaries that are discontinuous in depth (Marr and Poggio, 1976).

2 While truly transparent surfaces may occur infrequently in the nature, semi-transparency (fences, bushes, grass viewed by small creatures against the horizon, etc.) is a general phenomenon.

3 The epipolar lines are straight lines (on a planar projection surface) that match point by point. When the observer fixes a distant object these lines are horizontal and parallel, i.e. matching can be performed raster by raster. In general, their orientation depends on the mutual direction of gaze of the two eyes. It appears possible to compute the directions of the gaze (and thus the orientation of the epipolar lines) without any extra-axial information (Prazdny, 1983). If the epipolar lines are known, the matching problem is essentially one-dimensional.
same 3D object be similar, i.e. neighbouring image points corresponding to the same object should have nearly the same disparities. This suggests that the principal disambiguation mechanism should be facilitation due to disparity similarity. Dissimilar disparities should not inhibit each other because in a transparency situation, a disparity may be surrounded by a set of features corresponding to other surfaces. Two disparities are either similar, in which case they facilitate each other because they possibly contain information about the same surface, or dissimilar in which case they are informationally orthogonal, and should not interact at all because they potentially carry information about different surfaces.

Some Computational Considerations

In order to translate the coherence principle into a working algorithm one has to explicitly specify what is meant by disparity similarity, i.e. one has to develop a measure quantifying the similarity between neighbouring disparities. As a first approximation, we want a simple scalar function capturing the following three requirements:

1. The disparity similarity function should be inversely proportional to the difference of disparities of interacting points.
2. More distant points should exert less influence while nearby matches should have more disambiguating power.
3. The more distant the two interacting points are the less seriously should their disparity difference be considered because of the inherent uncertainty: steeply-sloped surfaces will generate large disparity differences which should nevertheless contribute to disambiguation. For large separations one should probably expect a (nearly) flat support function, i.e. all disparity differences should have the same influence. For small feature distances exactly the opposite should be the case: the probability of a large disparity difference is zero in the limit, i.e. the distribution function should peak at the centre.

One similarity function capturing all of these requirements is the familiar Gaussian distribution function

$$s(i, j) = \frac{1}{c|i-j|^2} e^{-\frac{|d_i - d_j|^2}{2c^2|i-j|^2}}. \quad (1)$$

Here, $s(i, j)$ expresses the amount of support disparity $d_i$ at a retinal point $i$ (a vector) receives from disparity $d_j$ at another point $j$, and $|i-j|$ is the distance between the two retinal locations (scaling constant $c$ is explained below).

There are several important points to be made about this similarity function.

a) The disparity difference in the exponent of the gaussian weighting function is scaled by the spatial separation of the two interacting points. This scaling means that the spread of the gaussian (controlled by $c|i-j|$) will be greater for widely separated points. Thus, distant point with a large disparity will be contributing some support whereas nearby points with the same large disparity will give little or no support.

b) Because of the way $c|i-j|$ controls the whole shape of the similarity function, when spatial separation is small the size of the weighting from nearby matches is greater.

c) The square root of the non-constant term of the exponent of the gaussian weighting function, $\frac{|d_i - d_j|}{|i-j|}$, is related to the disparity gradient (Burt and Julesz, 1980). In fact, this terms is the disparity gradient measured on a monocular (as opposed to cyclopean) visual manifold. Increasing the disparity difference increases the gradient and decreases the magnitude of mutual support between the two disparities. The major difference between this formulation and the notion of limiting disparity gradient proposed by Burt and Julesz (1980) is that their formulation implies the existence of inhibition between two interacting points while in our formulation two greatly different disparities simply do not interact at all. Interestingly, it has been found recently that the value of the limiting disparity gradient is a function of feature similarity, i.e. it cannot be defined on purely geometrical basis (Prazdny, 1985). More dissimilar features allow larger disparity gradients.

d) Maximal support is obtained when the disparity difference is zero independently of the distance between the interacting features. This means that an algorithm based on such a weighting function is slightly biased in favour of frontoparallel surfaces. This should not, however, be regarded as a weakness. The algorithm described below can successfully process steeply-sloped surfaces simply because similar disparities provide greater support than dissimilar ones. The bias towards the frontoparallel planes can also be defended on probabilistic grounds. It has been shown (Arnold and Binford, 1980) that, because of foreshortening, surfaces with steep depth gradients occupy only a small portion of most images. Most of the area of the image is covered by surfaces with normals to the viewing direction.

In the following, we tacitly assume that the stereograms were obtained by an imaging system in which the separation between the two viewpoints is small.

---

4 Recently, Pollard, Mayhew, and Frisby (1984, personal communication) formulated an algorithm based on the concept of the disparity gradient.
relative to the viewing distance. These requirements are met by all biological visual systems because the interocular separation is small relative to distances for which the binocular stereo mechanism is useful.

The Algorithm

The coherence constraint and the reasoning pertaining to the disparity interaction can be translated into a set of explicit rules: an algorithm. A computer program implementing the algorithm is outlined below. Basically, the program proceeds in two stages. First, we find all potential disparities (i.e., allowable correspondences between a point in the left and right image) for each point in the left image. Associated with each possible disparity is an "activity cell" whose value indicates the amount of support the particular disparity receives from its neighbors. Next, the disparities are allowed to influence each other. This interaction is a simple facilitation expressed by Eq. (1). More precisely, suppose that a left image feature point at location \( i \) has a set of possible disparities \( D_i \), and that we are interested in the amount of support a particular disparity \( d_i \in D_i \) receives from the feature point \( j \) (with possible disparities \( D_j \)). The algorithm locates, among the disparities in the set \( D_j \), that disparity \( d_i \) for which the absolute value of the difference \( \delta d = |d_i - d_j| \) is the minimum, and increments the "activity cell" associated with \( d_i \) at \( i \) by \( s(i,j) \) found by evaluating Eq. (1). Observe that in searching for the support for \( d_i \) among disparities \( D_j \), we consider only the best possible support that \( d_i \) can get there, i.e., only the support from \( d_j \) is considered! This is a non-linear step which not only ignores the irrelevant information but also leads to significant computational savings.

After the support for all possible disparities at a given point has been determined in this way the disparity with the largest support (the highest value in the associated "activity cell") is chosen as the most likely disparity at that point. The procedure as described above is run for the left and right image in parallel and all points for which the left and right disparities differ are marked as ambiguous (the decision about the disparity at such places is postponed and left for the later processes to decide).

It is possible to implement the coherence principle in another way, using strictly local interactions. The "global" support function [Eq. (1)] can be approximated by one that results from iteration of local finite difference equation. In this way, a tradeoff can be made between a large number of connections and a number of iterations necessary to propagate information (see also Szelisky and Hinton, 1984).

Algorithm's Performance

The performance of the algorithm depends on the value of \( c \) in Eq. (1). This is the only "free variable" in the algorithm. If \( c \) is small, the algorithm works very well for fronto-parallel surfaces but performs rather

---

Fig. 2. A noisy stereogram obtained by digitizing a stereogram published in Julesz (1971) and a disparity field detected by the algorithm using this stereogram as input. Intensity codes disparity magnitude. The amount of incorrectly detected disparities is small (300 points out of 19,000 were assigned an incorrect disparity)

---

5 In our algorithm, this decision takes the form of "interpolation". We simply look around the given point, make a histogram of the neighbouring disparities found by the detection process (using a rather narrow point spread function) and choose the disparity closest to the histogram peak
poorly for other surface orientations. We have found empirically that a value between 0.55 and 0.85 gives a good all round performance for a wide range of surface orientations.

The algorithm was tested on random-dot stereograms and some natural images. Random dot stereograms are a suitable vehicle for testing the performance of the algorithm because in these displays it is easy to obtain positional information needed for matching. All stereograms were pre-processed first to obtain edge information. Because the random-dot stereograms contain only two brightness levels (black and white) extracting edge information amounts to detecting the dark/light transitions and their signs. The program takes this information as input and matches only features of the same contrast polarity (edges of the same sign). The algorithm was tested on a broad range of such stimuli. It performs with accuracy close to 100% in stereograms portraying opaque surfaces. The accuracy decreases when transparent surfaces are present (about 75% of points were matched correctly). We believe that this is a rather good performance, possibly at the level achievable by human vision. Figure 2 illustrates the algorithm’s performance on a noisy stereo image pair obtained by digitizing and thresholding a random-dot stereogram published by Julesz (1971). Figure 3 shows algorithm’s performance on a natural image. The two half-images were first convolved with an “edge” operator to obtain the

6 This stereogram is digitized directly from Poggio (1984, pp. 107)
necessary positional information for matching (Fig. 3b). Only the position of the "edges" and their polarity were used as features for matching. Figure 3c shows detected disparities. The results of experiments with the algorithm using stereograms portraying opaque surfaces compare favourably with the result obtained using other existing stereomatching algorithms, e.g. the algorithm developed by Marr and Poggio (1979) and implemented by Grimson (1981).7

Discussion and Conclusions

The algorithm exhibits several important and interesting properties. It is, unlike previous methods, a non-iterative, parallel and local process. The interaction mechanism itself is a form of generalized correlation which allows for (local) deformations and belongs to a class of methods based on local consensus (Julesz, 1971; Stevens, 1978; Ballard et al., 1983). In some sense, the behaviour of the algorithm can be described as a distributed (global) maximization of the total support for (unique) local matches.

Another important point is that the coherence concept used as a guiding principle in our analysis directly subsumes several matching and disambiguation rules commonly used as explicit matching predicates to solve the correspondence problem. Figural continuity and edge connectivity (Mayhew and Frisby, 1981; Baker and Binford, 1981; Mayhew, 1983) or collinearity (Mayhew and Frisby, 1980) are special instances of the coherence principle because spatially continuous surface markings directly imply disparity similarity. Human visual system apparently does not use edge connectivity as a constraint. Recently, Krol and Van Grind (1980) demonstrated that a retinally continuous bar can be seen as two bars separated in depth. In their experiments with the double nail illusion (in which two pins are presented one behind the other in the mid-sagital plane) they observed that a part of a pin appears to float above, behind and between two other pins while both pins project into two continuous "bars" on the retina, i.e. a retinally continuous edge is phenomenally split into two parts in depth (Krol and Van Grind, 1980, Fig. 4).

The form of gathering supporting evidence from the neighborhood can be extended by measuring local image intensity gradients and "deforming" the shape of the support neighbourhood (currently a horizontally extended rectangle) according to "predominant" (e.g. modal) gradient orientation in the neighbourhood.

The idea is that the disparity of features along the direction perpendicular to the gradient varies potentially less than disparity of features along the local (luminance) gradient direction. This amounts to introducing a bias into local interactions in favour of local continuity. Observe that this is different from postulating a separate orientationally tuned nonlinear grouping process (Mayhew and Frisby, 1980) which explicitly uses the edge connectivity as a matching constraint.

An interesting feature of our approach is the absence of any explicit inhibitory connections (both in the spatial and disparity domain) in the algorithm, sometimes regarded as essential to disparity detection (Sperling, 1970; Dev, 1975; Nelson, 1975). The coherence principle does not penalize dissimilarities by inhibiting deviations. We believe that this is the major reason why the algorithm copes naturally and rather successfully with transparent surfaces.

The algorithm is not intended, in its present form, to be a performance model of the fusional process although its performance is demonstrably good. It does not address, for example, the issues of multiple spatial scales and their interaction8 the type of matching primitives to be used, or the role and control of vergence eye movements.9 These considerations are, however, in some sense, orthogonal to the major issue addressed in this work: a fast matching mechanism that does not rely on artificial constraints and can successfully operate in a general environment. Our results demonstrate that useful disparity information can be obtained fast and reliably without a need to postulate inhibitory connections between detectors tuned to different disparities, a coarse-to-fine strategy requiring a sequence of fine vergence eye movements, or an elaborate set of matching and disambiguation rules. Instead of trying to escape the ambiguity problem the algorithm exploits regularities in the distribution of local ambiguities necessarily following from the cohesiveness of matter.

Acknowledgement. I wish to thank Dr. J. Marty Tenenbaum for useful comments and discussion.

---

7 The result in Fig. 3 can be directly compared with the result published by Poggio (1984, figure on p. 107) who used the same stereogram.

8 In one version of our algorithm, the multiple spatial scales are handled by assigning a slightly greater importance (weight) to matches at more coarse levels. Again, this introduces a bias into, rather than imposing an order on the course of the matching process.

9 It is possible that vergence eye movements may represent the main method for dealing with transparent surfaces. Such accounts hold that transparent surfaces are perceived because the visual system fixates, in turn, each surface independently. The idea is that first the features belonging to one surface are fused and thereafter held "locked" together while a new vergence eye movement enables the next surface, with a different disparity, to be fused.
References


Received: January 29, 1985

Dr. K. Prazdny
Artificial Intelligence Laboratory
Schumacher Palo Alto
Research Center
3340 Hillview Avenue
Palo Alto
CA 94304
USA