Dynamic Edge Warping: An Experimental System for Recovering Disparity Maps in Weakly Constrained Systems

K. L. Boyer, Member, IEEE, D. M. Wuescher, and S. Sarkar, Student Member, IEEE

Abstract—A new technique called dynamic edge warping (DEW) for recovering reasonably accurate disparity maps from uncalibrated stereo image pairs, is presented. That is, no precise knowledge of the epipolar camera geometry is assumed. The technique is embedded in a system including structural stereopsis on the front end and robust estimation in digital photogrammetry on the other for the purpose of self-calibrating stereo image pairs. Once the relative camera orientation is known, the epipolar geometry is computed and the system may use this information to refine its representation of the object space. Such a system will find application in the autonomous extraction of terrain maps from stereo aerial photographs, for which camera position and orientation are unknown a priori, and for on-line autonomous calibration maintenance for robotic vision applications, in which the cameras are subject to vibration and other physical disturbances after calibration. This work thus forms a component of an intelligent system that begins with a pair of images and, having only vague knowledge of the conditions under which they were acquired, produces an accurate, dense, relative depth map. The resulting disparity map may also be used directly in some high-level applications involving qualitative scene analysis, spatial reasoning, and perceptual organization of the object space. The system as a whole substitutes high-level information and constraints for precise geometric knowledge in driving and constraining the early correspondence process.

I. INTRODUCTION

The perception of depth from passive stereopsis by computer is addressed in this paper. In this case, only the natural lighting of the scene is available and depth information is extracted from the photometric features of a stereo image pair. In constructing a visual stereo system, two problems are paramount: 1) Correspondence: The (relative, at least) position and orientation of each camera must be determined with respect to some world coordinate frame. The correspondence solution establishes the image plane parameters of the local triangulation, while calibration allows the results to be related to the real world. Additionally, a well-calibrated camera system is necessary for most stereo correspondence algorithms, since this provides geometric constraints that are used to direct the search processes. The primary difficulty associated with the correspondence problem is its immense combinatorics; much stereopsis research centers on the identification and enforcement of constraints to allow a suitable mapping to be recovered in reasonable time. Generally, these constraints derive from knowledge of the stereo camera geometry, or calibration.

Geometric constraints come in two flavors: camera-based and scene-based. We are specifically considering camera-based geometric constraints at the moment, those having to do with camera position and orientation. Scene-based geometric constraints have to do with hypotheses about the nature of the allowable surfaces in the scene. Virtually all stereo correspondence algorithms require the first, many also make use of the second.

The stereo correspondence problem has been pursued by many researchers [1]–[18]. These techniques generally operate at the pixel (point) based level with little aid or guidance from higher order image structure. To render the problem tractable and allow unambiguous solutions to be achieved, virtually all of these paradigms incorporate the epipolar geometry constraint. Under these conditions, the two image coordinate systems are assumed to differ in one coordinate only. That is, the camera axes are parallel, no rotations, and so on. Attaining this condition (or calibrating the deviation from it) requires substantial up-front effort.

Other techniques employ a so-called rubber mask or elastic matching approach [6]–[8]. In these techniques, the system matches the two images in some global sense, and then distorts one image locally to fit the second. Different types of distortions are assigned different costs. Although this technique can yield good results, it is computationally expensive, and some implementations require that the variance in depth throughout the image be minimal so that a good global match can be achieved with only a translation and rotation. This is ineffective for scenes having a large depth range, because a good global fit based on an affine transformation is impossible under those conditions.

In aerial photography the positions and orientations of the cameras are not known, even relative to one another,
a priori. Thus, it is necessary when considering stereo vision systems that employ a moving platform to ascertain (relative) camera position and orientation parameters from the stereo image data itself and not from special calibration scenes or fixtures. In other settings, such as robotics, it is advisable to check the calibration parameters periodically to monitor the effect of equipment vibration and physical disturbances that may occur over time. Robust estimation in analytical photogrammetry [9] may be applied to this calibration problem, but requires that many corresponding points be identified in the two images. The identification of those correspondences automatically by computer for an uncalibrated camera pair has been, heretofore, difficult.

This paper presents an important step in developing such a system. We introduce a procedure that we call dynamic edge warping (DEW) by analogy with the dynamic time warping (DTW) concept in speech recognition. With this technique, dense point matches are obtained along extended edge contours that have themselves been matched by structural stereopsis [10], [11]. Since neither of these techniques relies on a precise epipolarity condition, they form the core of a stereo system capable of “bootstrapping” itself from a very weakly constrained starting point.

Our approach to constraining the problem is to begin by matching higher level primitives. This reduces the total number of primitives to be matched, and allows us to attach more descriptive information to each. Matching structural descriptions of the two images, built in terms of these primitives and their spatial relationships, offers a highly robust primitive correspondence system. Boyer and Kak [10] and Sotak and Boyer [11] followed this route. Boyer and Kak presented the initial conceptual and theoretical groundwork for structural stereopsis and demonstrated the concept on structural descriptors derived from radial-valued skeletons of binary images. Sotak and Boyer subsequently presented a structural stereo matcher using extended chains of Laplacian-of-Gaussian zero crossings as structural primitives, providing the first application of structural stereopsis to gray-level images.

Unfortunately, structural matching of extended primitives, such as chains of edge points, does not yield pixel-level disparity information directly. Instead, this matching result decomposes the overall 2-D point matching problem across the image pair into a set of 1-D point matching problems across corresponding, or conjugate, contour segment pairs. By this means we can substitute other constraints for epipolarity and thus solve the correspondence problem when the camera geometry is, for all practical purposes, unknown. The point correspondences thus recovered are then used to compute the relative pose of the camera pair and so compute the epipolar geometry which, in turn, allows us to compute a depth map.

We note philosophical similarity to recent work by Tanaka and Kak [12] that attempts to match higher-order primitives (long straight lines in the reported implementation) and thereby generate disparity hypotheses to drive a fine-channel Marr–Poggio–Grimson (MPG) matcher. Their approach is rule-based, rather than graph-theoretic, and still relies on the epipolarity condition as a driving constraint.

We are not the first to consider the problem of self-calibrating stereo. Hannah’s Bootstrap Stereo [13], [14] is designed with goals much the same as ours, but we approach the problem very differently. We will have more to say on this in Section V. Gennery [15] describes a system which estimates relative camera orientation parameters as it progresses, using Bayesian decision theory to accept or reject matches between image structures, each characterized by a measured feature vector. The reader is also referred to related work by Perlman and McKeown [16] which begins to unify registration and matching in aerial imagery. We also note work on self-calibration by Takahashi and Tomita [17] which begins with matched boundary representations.

We would be remiss, of course, if we failed to mention that Mayhew and Frisby [18] pioneered the notion of an extended (nonpoint) stereopsis primitive in a computational theory. Theirs was the concept of “figural continuity,” later incorporated by Grimson [19], as well. Nonetheless, these efforts still require knowledge of the epipolarity condition and use the figural continuity constraint in only a limited sense. Our top down approach, beginning with the structural matching of constant curvature primitives, then proceeding to DEW, the topic of this paper, thus represents a significant departure from previous work.

In Section II we describe the problem setting, present the rationale for our algorithm, and give an overview. Section III presents a detailed description of the algorithm. Section IV presents experimental results including structural matching and DEW point matching. In Section V we offer some closing comments and observations, while the Appendix describes the computation of epipolarity and compares calibration from DEW correspondences with calibration from hand-selected conjugate points.

II. ALGORITHM RATIONALE AND OVERVIEW

The goal is to recover stereo correspondence between image pairs for which we have essentially no advance knowledge of the cameras’ position and orientation, i.e., the system is uncalibrated. Once sufficient numbers of conjugate points are identified, methods from analytical photogrammetry are used to calibrate the stereo setup. The basic system design assumptions are the following.

- We know which image is left and which is right.
- We know that the epipolar plane is more or less horizontal, within, say ±20° in each image. (As we will show in the experimental results, even this constraint can be seriously violated without ill effect.)
- The depth of field is reasonable, and about the same for the common fields-of-view.
- We have some limits on the degree of distortion between the two views, perhaps statistically characterized, for the class of images under consideration. This is used in structural stereopsis as described in [10].
- The distance to the scene content is about the same for each camera. That is, scale variations between the two views will not be extreme except, perhaps, locally. This is related to, but slightly different than, the previous item.
- The process is conducted off-line (calibration and subsequent object space description in aerial photography) or in the background (calibration maintenance in robotic vision). In either event, real-time performance is not mandatory, but reasonable efficiency is.

A. The Problem Setting

The basic premise of the system is that the Gestalt action of structural stereo allows robust solutions to the correspon-
dence problem at an extended primitive level. These solutions then seed a detailed point matcher to recover precise point correspondences along the matched contour segments.

In this system, the effect of structural stereopsis is to decompose a very large, weakly constrained, imagewise correspondence problem into a set of relatively small, well-constrained, local correspondence problems.

In our experiments, we achieve satisfactory performance by applying the system at a single level, but for very high resolution, detail-rich aerial images (4096 x 4096, for example) it may be prudent to construct a multiscale matcher for computational efficiency. In the more comfortable world of 512 x 512 images and smaller, the scale space concept appears to buy us little. Fig. 1 presents a block diagram of an autonomous camera calibration system.

- The process begins at the top with the application of the LoG (or some other edge operator) to each image. For these experiments we used the fast LoG implementation described by Sotak and Boyer [20] (see also Chen et al. [21]).

- The resulting contours are partitioned into segments of constant curvature. Although circular arcs are not strictly viewpoint-invariant, even cameras converged at up to 20 degrees give similar contour decompositions.

- The constant curvature segments are then used as primitives in a Parametric Structural Description, as defined by Boyer and Kak, for each image. Primitive attributes include curvature, contrast, polarity, mean-orientation, and length. Relations include pairwise orientation (of the line joining primitive centroids) and pairwise distance (between centroids).

- Given the interprimitive mapping function produced by the structural stereo module, DEW, the primary subject of this paper, is invoked for each matched contour segment pair to produce a vector-valued disparity map. A simple form of dynamic programming and linear prediction, subject to photometric constraints, produces a dense set of point-to-point matches.

- The point matches produced by DEW are then input to an orientation module which computes the relative orientation of the two cameras using robust estimation techniques. The system has now autonomously extracted the epipolar geometry of the image pair. This is sufficient to construct a relative depth map of the scene, control points being necessary to tie the result to an absolute coordinate system. This result can also verify that the camera system is still in calibration from an initial setup.

- Once the epipolar constraint is computed from the relative orientation, the system may return to the image pair and the structural interprimitive mapping function to compute a refined depth map which can, in turn, be used iteratively to refine the epipolar constraint, and so on. If control points can be autonomously extracted from the images, absolute orientation may be inferred and an absolute representation of object space computed.

B. Sources of Information and Constraints

The DEW algorithm is driven by the following criteria and observations:

Local Photometric Consistency: We expect that pixels taken either from the two contours, or around them, should be similar in some respect in a neighborhood around conjugate points. Experimentally, we determined that a “one-sided” measure of contrast (below) provided the most reliable photometric information source.

Global Consistency: The extraction of our structural primitives virtually assures us that all of the image points in a given contour segment arise from the edge of a single opaque object or marking in the scene; any other event giving rise to a constant curvature edge segment would be an extreme accident of object(s) and camera(s) alignment. Thus, there are two aspects to global consistency.

Ordering: The correct point-to-point mapping function along a contour pair should be such that basic spatial relationships are preserved. In the case of DEW, the spatial relationship of interest is ordering; the mapping should preserve ordering relationships among the contour points. This constraint derives from a simple observation: in dealing with opaque objects, the ordering of object points must be, for a given object, consistent in the two views. Tanaka and Kak [12] have termed this condition “monotonicity of rendition.” Several implementations have applied this constraint along epipolar lines in a dynamic programming framework, for example [22]-[24].

Disparity smoothness: Since our primitives are assumed to arise from a single (3-D) object edge, we expect no large jumps in disparity (which would correspond to depth discontinuities) along the contour segment. Notice that we are not constraining the behavior of the depth map across the seg-
ment. Neither are we assuming any particular form for the local disparity behavior, only that it should vary smoothly.

We expect an edge to undergo a change of curvature (perhaps highly localized, like a corner) anywhere there is a discontinuity in depth, or the rate of change of depth, along its length. When it does, the contour decomposition algorithm will break the contour at that point. Thus, the contour segment pairs as submitted to the DEW algorithm should exhibit no discontinuities in either disparity or its first positional derivative. This form of geometric constraint is far less restrictive and easier both to defend and to verify than hypotheses such as local surface planarity or quadratic behavior [25], [26]. Lowe [27] captures this notion in his “non-accidentalness” constraint. It is highly unlikely in a single image, and virtually impossible in each of two distinct images, that two unrelated scene edges will align in such a manner as to project to a single circular arc.

Before developing the DEW algorithm, we should clarify a few key points. Edges not in the common field of view of the two cameras will not be considered by the algorithm because the structural stereo matcher [10] is quite reliable in not assigning false correspondences. This is not magic, it is a consequence of the richness of the structural descriptions. We again emphasize that structural stereo does not require the epipolarity condition. That would, of course, be self-defeating. All that is necessary is a reasonable characterization of the expected degree of image to image structural distortion, primarily in terms of angles and lengths.

C. General Structure of the Algorithm

The algorithm divides naturally into three parts.

Initial Point: The algorithm first recovers an initial point match of high confidence between the two contours an “island of photometric reliability.” Both global and local information sources are tapped in the selection of this match.

Dynamic Programming: The point-to-point mapping function is extended in both directions by dynamic programming. Global and local consistency criteria derived from photometric and geometric considerations are used together with linear prediction to drive the process.

Decision: The algorithm actually constructs two mappings, each driven by a different photometric constraint. As we will explain below, in general one of these mappings will be “correct” while the other may be erroneous, depending on the physical conditions producing the photometric edge. It is easy to select the better of the two.

D. The Basic Matching Criterion: One-Sided Contrast

Because structural stereopsis exhausts the geometric information available, we turn to photometric measures. The principle of photometric invariance in stereo vision is hardly new, simple photometric correlation approaches being among the first attempts at computer stereopsis. Hannah’s work [13], [14] represents the most notable, and probably the most sophisticated, of these. Our difficulty is applying the concept of photometric invariance precisely in those neighborhoods where it is unlikely to hold: around edges. Photometric invariance cannot be expected to hold near occlusive boundaries and, of course, many of our LoG contour segments will correspond to occlusions. Clearly, gray level along the edge and total contrast across the edge will not prove reliable.

The first suffers from trying to make a measurement right on the edge; the second fails for viewing geometries with different object surfaces to one side or the other, including virtually all occlusive edges and many others.

Instead, we develop a simple notion we call “one-sided contrast,” which circumvents these problems. We define two quantities which roughly characterize the LoG response: mag+ and mag− are the magnitudes of the first extrema to either side of the zero crossing. The sum of these two may be taken as a measure of the edge’s total contrast at the point in question. However, as noted previously, this is not a reliable photometric quantity for inferring contour correspondences. Nevertheless, the behavior of whichever of these quantities falls on the foreground side of the edge is quite consistent between the two images, as we shall see.

Fig. 2 presents the left and right image mag + and mag − plots for a conjugate contour pair. It is relatively clear that the mag + plots are similar in overall structure, while the mag − plots, although exhibiting some similarities, are certainly much more weakly correlated. Notice that it is not possible to simply shift one contour’s plot over the other to align them; there is a significant amount of viewpoint induced foreshortening between these two images. The edge in question (of a foam block, shown later) extends for roughly 185 pixels in the left image and about 225 pixels in the right.

The correct point-to-point is not one-to-one everywhere along the contour pair.

Since only one side of the edge corresponds to the same physical surface in both images, photometric invariance will only hold to one side. The reader may be concerned at this point that our quantities mag+ and mag−, either of which we term the one-sided contrast, are not independently determined by the two sides of the edge. In fact, if we assume the simple step edge model, these two quantities are identical in magnitude. However, the key observation is that edges are not (generally) purely odd functions of position to either side of the inflection point, which for the LoG is more or less associated with the zero crossing. The LoG response is a purely odd function of position with respect to its zero crossing point if and only if the edge function is odd.

To show that this claim is reasonable, we will offer a simple analysis that suffices for our purposes. For more thorough analyses of aspects, such as positional bias and localization, of operators based on zero crossings of second derivatives, the reader is referred to [28], [29]. Neither of these considers the effect of edge asymmetry on one-sided contrast measures.

Consider two 1-D edges, say e1(x) and e2(x). Suppose the first edge, e1 is purely odd, meaning e1(−x) = −e1(x) for all x, but the second, e2, does not meet this condition. Now, of course, the (1-D) LoG is even:

\[ g^+(x) = \frac{-1}{\sigma^2} \left( 1 - \frac{x^2}{\sigma^2} \right) e^{-x^2/2\sigma^2}. \]

Further, we recall the following property of the Fourier transform [30]:

\[ \mathcal{F} \{ f_s(t) \} = F_s(w) \quad \mathcal{F} \{ f_o(t) \} = F_o(w) \]

where \( f_s(t) \) and \( f_o(t) \) denote purely even and odd functions of time (or space), respectively, and \( F_s(w) \) and \( F_o(w) \) denote purely even and odd functions of (spatial) frequency, respec-
tively and $\mathcal{F}(\bullet)$ denotes the Fourier transform. Let $E(\omega)$ and $LG(\omega)$ represent the Fourier transforms of the edge signature and the LoG, respectively. Then the transform of the filter response, $O(\omega)$, is:

$$O(\omega) = E(\omega)LG(\omega).$$

(3)

Since the LoG is purely even in space, it is also purely even in frequency ($LG(\omega) = LG(-\omega)$ for all $\omega$). Likewise, the edge signature will be purely odd in frequency only if it is purely odd in space. So in the case of $e_s(x)$, the output will be purely odd in frequency (the product of an odd function and an even function is odd) and, therefore, will be purely odd in space. That is, $O(-\omega) = -O(\omega)$ for all $\omega$ and $o(-x) = -o(x)$ for all $x$. If the edge signature is not purely odd, such as $e_s(x)$, then it must contain an even part. If so, the edge signature will not be purely odd in frequency and the filter response will contain an even part in both frequency and, therefore, space. Thus, for this case, $o(-x) = -o(x)$ for all $x$.

Thus, for real edges we may expect the quantities $\text{mag}^+$ and $\text{mag}^-$ to behave differently. In particular, whichever of these two appears to the foreground side of the edge may reasonably be expected to behave similarly in the two views, while the other (if associated with a background region, especially one which is photometrically busy and under significant occlusion) may be expected to differ. Since $\text{mag}^+$ and $\text{mag}^-$ are predominantly functions of the image behavior in the immediate vicinity of the edge contour to their respective sides, and since at least one side of the edge is the same in both views, at least one of the two one-sided contrast measures should prove suitable to drive the point matching process. The two one-sided contrast functions independently construct two point-to-point mappings and a global consistency measure (below) selects the better of the two.

We are not the first to characterize edges by quantities measured independently to either side. Baker and Binford [22] used edge contrast together with image slope and intensity to either side to define the photometric properties of an edge.

III. DETAILS OF THE DYNAMIC EDGE WARPING ALGORITHM

The following sections use the segments above as an example contour pair throughout. (We will see the gray level and edge images presently.) $I_i$ refers to the pixel at the $i$th
location along the left contour, while \( r_j \) denotes the \( j \)th pixel from the right contour. The \( c_j(i) \) denotes the value of the (one-sided) contrast plot at the \( i \)th left contour pixel, while \( c_j(i) \) is defined similarly. Each pixel, say \( l_i \), resides at some physical location in its image, \((x_l(i), y_l(i))\). The disparity vector between any two contour points is indexed by the pair of pixel indices: \( d(i,j) = (x_l(i) - x_l(j), y_l(i) - y_l(j))^T \), where \( T \) denotes transpose.

A. The Initial Match: “Islands of Photometric Reliability”

The first problem in warping one contour onto the other is that of getting started. The decomposition of the original Log Gabor contour is quite obvious but certain not perfect. Adaptive effects and other view-to-view distortions are present. For these reasons, we cannot assume that endpoints of conjugate contours are necessarily conjugate points themselves; this is actually quite rare. So we first select a pair of points (one from each contour) for which there is strong evidence that they are conjugates. We then extend the point-to-point matching outward in both directions from the starting point. Referring to Fig. 2(a) and (b), “prominent” points such as \( c_{1(177)} \), \( c_{1(222)} \), and \( c_{1(2)} \) clearly provide the best opportunities to “get started.”

Given these observations, the next question to address is that of embedding the “prominent point” idea into a useful, reliable algorithm. The essence of the problem is demonstrated in our example. What tells us that \( l_{177} \) is, first of all, likely to correspond only to either \( r_{22} \) or \( r_{222} \)? And, moreover, how is it that no one considers the correct matching of \( l_{177} \) to be \( r_{22}^2 \)? It is not difficult to postulate two categories of information to drive this process. The first is local, a similarity measure based on the one-sided contrast around the candidate match points. This is the information in \( c_{1(i)} \), identifying \( r_{22} \) and \( r_{222} \) as the only reasonable candidates to match \( l_{177} \). Additionally, we should select points such that the similarity is remarkable. If the two contrast plots are constant-valued over some significant range, points taken from this range will exhibit exceptional similarity, but they will not be useful because they are not distinctive.

The second category is global information that measures match consistency from a relational standpoint (i.e., ordering) and the smoothness of the warping function. This directs us to associate \( l_{177} \) with \( r_{22} \) and not \( r_{222} \). Notice that these two classes of information are consistent with those which arise in structural matching [10], [31] and other incarnations of the consistent labeling problem [32], [33], i.e., local measurements on the objects to be labeled and relations over the set of allowable labels.

To apply these notions, we seek points of photometric reliability along each contour segment, as well as some indication as to the most likely value of disparity (as a vector) to be encountered. We do this as follows.

1) For each segment of the conjugate pair, identify several relative extrema of the contrast plot having the largest second-derivative magnitudes. (See Fig. 3 for the points selected for our running example.) We use 1/10 the contour segment length in pixels for a reasonable number of options without undue computational hardship; the precise number is not critical. To control noise, the first derivative is computed by convolving the contrast function with the first derivative of a Gaussian, \( \sigma = 1 \). The second derivative at each point is taken to be the difference in the first derivative to either side.

2) For each possible pair of local maxima, one from each contour, chosen in the previous step, compute the corresponding disparity vector, \( d(i,j) \). Do the same for each pair of local minima.

3) For each selected pair of maxima and minima, compute the vote strength:

\[
v(i,j) = \frac{\min \{ |c_l(i)|, |c_l(j)| \}}{\sqrt{\max \{ |c_l(i)|, |c_l(j)| \}}} \tag{4}
\]

This simple weighting assigns more importance to points having similar second derivatives (the \( c^2 \)'s) of the contrast function, especially when the second derivatives are large. Thus very steep peaks and valleys with similar contrast shape will have large vote strengths.

4) This information is then assembled in a two dimensional weighted disparity histogram. Each pair of voting extrema, \((l_{1(i)}, r_{2(j)})\), contributes \( v(i,j) \) to the quantity at location \( d(i,j) \). For contour segments of reasonable length, we expect the dominant or average disparity along the segment receives a plurality of votes due to the coherence of the correct pairings while the incorrect pairings produce votes dispersed randomly over the histogram.

5) For short contour segment pairs, the disparity histogram may be quite sparse, making identification of the major mode unreliable. To ameliorate this problem, we confirm the peak value of the disparity histogram by comparing it to approximate disparity data gathered from across the image pair. The average and standard deviation of the disparity between the centroids of matched structural primitives are used to do this. If the peak histogram disparity lies within one standard deviation of the average centroid disparity, we accept the peak histogram disparity as the most likely disparity for the pair of segments in question. If not, we then compare the peak disparity with the centroid disparity of just the segment pair in question. If the segment centroid disparity is closer to the average centroid disparity than the peak histogram disparity, we reject the peak histogram disparity and use the segment centroid disparity as the most likely disparity for the pair in question.

6) Once we have identified the most likely disparity for the contour pair, we select that pair of photometrically reliable extrema (as above) whose disparity is closest to the estimated most likely value. In the event of a tie, we select the pair having the largest vote strength, \( v(i,j) \). In the very unlikely event of a tie in the vote strength as well as in the most likely disparity, the first one in the list is chosen arbitrarily; this has, in fact, never happened.

Using either the peak histogram disparity or the segment centroid disparity alone will usually work well in estimating the correct value and selecting a suitable starting point.

3Our definition of the centroid is a bit cavalier. We simply take it to be the midpoint of the line segment joining the endpoints of the contour segment.
However, as we indicated previously, short contour segments tend to have sparse disparity histograms, weakening the reliability of that approach for those cases. Yet the centroids of short segments must lie very close to the contour trajectory. Meanwhile, although long contour segments can have centroids displaced from the contour trajectory by several pixels, rendering the segment centroid disparity suspect, the disparity histogram for these contours is not sparse and so can be trusted. Thus, each estimator performs well when the other does not, and there is a significant region of overlap where both work well. The system thus builds consensus to ensure that we use the most reliable information available in predicting the most likely disparity.

In the case of our continuing example, the initial match point is \((l_{15}, r_{15})\), as circled on Fig. 3. Although this point happens to be nearly the centroid, this is just coincidence. The reader may wonder why this pair was chosen over the seemingly more obvious \((l_{115}, r_{253})\). The answer lies in the algorithm’s use of disparity information. The most obvious pairing based on photometric information alone lies away from the most likely disparity by several pixels while the pairing chosen lies very close to it. This does not prevent the algorithm from eventually pairing \(l_{177}\) and \(r_{222}\) as conjugates; all that’s been decided thus far is that \(l_{22}\) and \(r_{115}\) are certainly conjugates and that they form the most reliable place to begin.

### B. Extending the Mapping: Dynamic Programming

From the initial point, we use a simple form of dynamic programming to extend the mapping outward in both directions. We select the next available pixel from, say, the left image contour and match it to one of five candidate match points in the right. The left image contour position is not necessarily always the independent variable, with the right image contour position the dependent variable, as the mapping is constructed. For programming reasons, the roles of dependent and independent variable may occasionally be interchanged as the mapping is extended; the details are uninteresting and are omitted.

**The Basic Parameters:** For discussion, let the left contour pixel position be the independent variable and the right contour pixel the dependent variable. Let an entry in the mapping function be denoted; \((l_i, r_j)\), where \(i\) is the left contour pixel index and \(j\) is the corresponding right contour pixel index. From an initial match of, say, \((l_{10}, r_{15})\), the dynamic programming procedure first builds the mapping function in one direction by selecting left contour pixel 16 and comparing it with right contour pixels 14 through 18. (The comparison mechanism is described below.) The corresponding to left pixel 16 will be selected from among these five candidates. In the other direction, left pixel 14 is compared to right pixels 14 through 18, and so on. Examining, at each step, the most recently used dependent pixel and the next four limits the edge warping function to a local rate of expansion (compression) of 4:1. This is certainly very generous; we do not expect to see foreshortening ratios as high as 4:1 over extended regions, since the structural matcher would not assign correspondence under such conditions. We let the process look that far ahead to allow for locally severe foreshortening near corners, for example.

**Local Photometric Fit:** Local photometric similarity in the form of one-sided contrast drives the dynamic programming portion of the DEW algorithm. We construct two warping functions, one for each “side,” and then select the better of the two as described as follows. Given a measure of one-sided contrast (either \(\text{mag}+\) or \(\text{mag}−\)) at each contour pixel, we compute the following similarity measure, called the local photometric fit, between the candidate match points:

\[
 f_{\text{phot}}(i,j) = \sum_{k=1}^{n} \left( m(l_{i+k}) - m(r_{j+k}) \right) + \beta \left( + |m(l_{i+k}) - m(r_{j+k}) + \beta| \right)
\]

where \(i, j\) are indices along the left and right contour segments, respectively, \(l_i, r_j\) are the candidate match, the local window size is \(2n + 1\), \(m(\cdot)\) is either \(\text{mag}+\) or \(\text{mag}−\), and \(\beta = m(l_i) - m(r_j)\).

The local photometric fit just sums the absolute differences in contrast over a small window of points equally spaced from the candidate match, adjusted for the difference at the candidate match. Although we expect the contrast shape to be relatively invariant around conjugate points, we cannot expect the contrast value to be similarly consistent. The effect of foreshortening between the two contour seg-
ments (compression or expansion in the mapping) is negligible over the small (5 pixel) windows.

Disparity Prediction and Derating of the Photometric Fit: In many cases the photometric data alone is insufficient to identify conjugate points accurately, for instance when the contrast is nearly constant over some extended region. We therefore impose a disparity smoothness constraint along the conjugate contour pair. As argued above, this form of geometric constraint is both quite defensible and easily enforced because we are considering two views and structurally-matched, constant-curvature contour segments. It is therefore reasonable to assume that a scene feature giving rise to a contour segment of constant curvature in both image spaces is free of discontinuities in depth. Incidentally, Lowe [27] captures a similar notion (for monocular views) in his "non-accidentalness" constraint. So we predict the disparity vector of the next point match and bias our next decision toward that value, especially when the photometric information gives no obvious result. In effect, the system implements a sort of "flywheel" process to carry the matching across regions having no distinctive photometric signature. When there is sufficient photometric evidence to drive the process, it does so. This also adjusts the "speed" of the flywheel. When the process enters a region which is photometrically ambiguous, it "glides" across according to the most recently available information from interesting regions.

For each candidate match we compute the local photometric fit, \( f_{\text{loc}} \), which is then derated (rendered less attractive) by a positive-valued weighting function. The derating process begins with a linear prediction of the disparity vector for the next match. The algorithm fits a straight line through the previous four disparity values and extrapolates the next. The selection of four previous values to predict the next disparity is an engineering compromise between stability and responsiveness; it's not especially critical. The farther the disparity of a candidate match is from the predicted value, the more its photometric fit will be derated. The shape of the derating function incorporates our observations concerning disparity smoothness, our confidence in the prediction, and other factors. Consider the following observations:

1) To make matches away from the predicted disparity less likely, the derating function should monotonically increase as the candidate disparity moves farther from the predicted disparity and should vanish at the predicted disparity.

2) The slope of the derating function should be proportional to the difference between the predicted and candidate disparities. Away from the predicted value, we require increased justification (as measured by \( f_{\text{loc}} \)) for a match. The derating function should therefore take the form of an exponential "potential well" centered on the predicted disparity vector.

3) There should be some overall method of weakening, or "diluting," the disparity smoothness constraint when our confidence in the prediction is low; i.e., there should be some control as to how fast the exponential rises. Our confidence in the prediction is low in two situations:
   - When the dynamic programming process is just beginning and there is insufficient data from which to predict the next value.
   - When the variance of the established disparities about the best fitting straight line is large, indicating that the rate of change of disparity is either locally nonlinear, or significant disturbances are present.

4) We should also accommodate a priori knowledge of the variation of the \( x \) and \( y \) components of disparity. For example, if we know that the images are roughly aligned horizontally, then we can assume relatively small variation in the \( y \) disparity; deviations in \( y \) disparity from the predicted value are less attractive than equal changes in \( x \) disparity. In that case, the potential well should not be circularly symmetric, but should increase more rapidly in the \( \pm y \) direction than in the \( \pm x \) direction.

Considering these observations, we choose a derating function of the form:

\[
 w(i,j) = \exp \left( a \sqrt{\frac{(x'_i - x'_j - \hat{x})^2}{\mu} + \frac{(y'_i - y'_j - \hat{y})^2}{\nu}} \right)
\]

(6)

where, \( i,j \) are left and right contour pixel indices, respectively, \( a \) is the prediction-confidence dilution factor described in item 3, \( x'_j - x'_i \) is the candidate \( x \) disparity, \( \hat{x} \) is the predicted \( x \) disparity, \( y'_i - y'_j \) is the candidate \( y \) disparity, \( \hat{y} \) is the predicted \( y \) disparity, \( \mu \) is a constant to account for a priori knowledge of the variation in \( x \) disparity, and \( \nu \) is a constant to account for a priori knowledge of the variation in \( y \) disparity.

Fig. 4 illustrates this function for \( a = 1, \mu = 1.0, \) and \( \nu = 0.4 \). The bottom of the well (where \( w(i,j) = 0.0 \)) is placed at the predicted disparity. Notice that deviations from the predicted \( y \) disparity are penalized more severely than equal deviations from the predicted \( x \) disparity, for this example.

The numerical measure used to select among the candidate matches at each step is given by the product of the local photometric fit and the derating function:

\[
 q(i,j) = w(i,j) \times f_{\text{loc}}(i,j).
\]

(7)
We emphasize that the prediction of the next match point and the derating of photometric fit are done in 2-D disparity space, not in terms of the pixel ordinal positions along the contours. The predicted next match point, among the five candidates, will be that having disparity (with respect to the next independent contour point) closest to the predicted value. The selected next match point will be that whose derated photometric fit is best (numerically lowest). The smoothness criterion is therefore conceived, defined, computed, and enforced in terms of disparity.

The variable $a$ reflects our confidence in the (linear) prediction of disparity. If the variance with respect to the best linear fit is relatively low, then our confidence in the prediction is high, and we therefore increase $a$, forcing the next match point to lie closer, in disparity space, to the prediction. If the variance with respect to the best fitting line is high, then we lack confidence in the prediction, $a$ is reduced, and the walls of the bowl are less steep. We also reduce $a$ when the process is just starting (less than four previous points).

Selecting the Parameter Values: Given the roles of parameters $a$, $\mu$, and $\nu$, we now describe how to set them.

a) The prediction-confidence parameter is set as follows:

$$a = \max\left\{1.0, \left(5 - 6\sigma_p^2\right)\right\}$$

where $\sigma_p^2$ represents the variance of the predictor samples with respect to the best-fitting straight line, except for the starting conditions. More precisely, once the process is running, $\sigma_p^2$ is the sample variance of the absolute value of the difference between the best linear disparity and the actual disparity values. Under startup conditions, $\sigma_p^2$ is set to 0.5, so $a = 2$, until the predictor has four previous points available. As $\sigma_p^2$ varies from 0 to 2/3, $a$ varies from 5 to 1, respectively.

b) The $a$ priori knowledge of $x$ disparity variation is set to the standard deviation of the segment centroid $x$ disparities from across the image pair as obtained in the structural matching step. That is, to a first approximation, we expect the standard deviation of pointwise $x$
of *ad hoc* parameters, that proves difficult without the precise geometric constraints which we are, in fact, trying to determine. We acknowledge that (8) is a bit *ad hoc*, but is based on a solid rationale. Once the form of the equation was determined, the specific numbers were experimentally set. However, the set of images used to produce the specific numbers in (8) ranged over a wide variety of scene content, camera types, and geometry.

Fig. 5(a) presents a plot of the mapping for our continuing example contour pair. In this example, both endpoints nearly match, both image contours are straight, and the physical edge (of a foam block) is straight. Therefore, we can also plot a very close approximation to the perfect point to point mapping as the straight line shown. Overall, the mapping is quite accurate. The reader should note that, for this example, there is a significant degree of foreshortening: about 180 left image pixels are dispersed over about 225 right image pixels for a mapping that is approximately 4:5 expansive, left to right. The mapping of \( f_{175} \) to \( r_{222} \) is also accomplished. Four pixels from each end of the left contour were left unmapped as end effects; the process terminates when it runs off the end of either contour.

Fig. 5(b) plots the \( x \) disparity along the left contour. Since in this case, disparity decreases with distance from the camera, the result is very much as we expect.

*Selecting the Final Mapping Function:* There appears to be no reliable means of selecting the foreground one-sided contrast function prior to mapping. But while a point to point mapping between two dissimilar contrast functions may exhibit reasonable smoothness, it can be expected to exhibit poor photometric fit, in general. On the other hand, a point to point mapping constructed between the “correct” pair of contrast functions should exhibit superior photometric fit, possibly even at the expense of some degree of smoothness. Therefore, we run the initial match selection and dynamic programming procedure using each contrast function independently for each conjugate contour segment pair. The final step in the algorithm selects the better of the two pointwise mapping functions for each contour segment pair as that which yields the better average value of \( f_{loc} \). The other mapping is discarded.

---

**Fig. 6.** Tabletop scene. (a) Left gray level image (b) Right gray level image (c) Left segmented contours (d) Right segmented contours.

**Fig. 7.** Tabletop scene: \( x \) disparity image.
IV. Experimental Results: Disparity Maps

In this section we present experimental results on three stereo pairs, a tabletop laboratory scene, a foam block intentionally arranged to produce significant foreshortening, and a pair of aerial images. Although unrectified, all three pairs of images were taken such that most of the disparity lies in the $x$ direction. The camera axes were somewhat convergent for the tabletop scene, perhaps by as much as $15^\circ$; the camera baseline was about 4 feet and the objects were placed about 4 to 5 ft from the baseline. The aerial photographs were convergent or divergent by probably no more than $3^\circ$, and the foam block image was also taken with camera axes nearly parallel. In the Appendix, we describe the computation of epipolarity and compare orientation using correspondences obtained with the DEW algorithm with that using hand-selected correspondences. Surprisingly, the large number of correspondences identified by the DEW algorithm allows it to outperform manual selection of conjugate pairs for calibration in some cases.

The foam block images are $500 \times 482$ pixels, all others are $250 \times 241$; we applied the LoG as implemented in [29] with $\sigma = 3$. Zero crossings were detected with a predicate based algorithm similar, but not identical, to that presented in [34]. Whole contours with shallow average zero crossing slope ($<0.65$) were deleted. The remaining contours were partitioned into constant curvature segments according to the algorithm in [35], [36]. Contour segments with average zero crossing slopes less than 1.1 were then deleted. This thresholding process is somewhat reminiscent of Canny’s hysteresis thresholding [37]. Incidentally, Clark [38] has recently presented a more reliable means of distinguishing physically significant contours from “ghosts.”

The segments were then structurally matched under fairly liberal criteria, the rationale being to submit some “challenges” to the point matcher. Our structural stereo matcher is nearly identical to that described in [10]. The matched contour segments were submitted, together with the average and standard deviation of the corresponding segment centroid disparities, to the point matching algorithm. In the results which follow, we plot the $x$ disparity of the corre-

---

Figure 8. Foam block. (a) Left gray level image (b) Right gray level image (c) Left segmented contours (d) Right segmented contours.

Figure 9. Foam block: $x$ disparity image.

---

Those that are closed or those that intersect the image boundary.
sponding points as assigned by the point matcher. Since the image pairs are at least roughly aligned horizontally, most of the (relative) depth information is conveyed by the $x$ disparity.

Figs. 6–11 show the gray level, zero crossing, and $x$ disparity images. Frankly, the disparity plots are somewhat difficult to interpret without a bit of practice; our images lack sufficient resolution to render the disparity plots clearly. Unfortunately, this seems to be the most direct means of visually presenting the data. If one examines the tabletop disparity image (or its numerical data) carefully it can be seen that the disparities corresponding to the jar are smaller than those of the cup or screwdriver, the jar averaging around 55 pixels in disparity while the cup and screwdriver average around 60. This coincides with the physical arrangement of the cameras and the objects in the scene.

The foam block is included as a test case to verify that the point matcher does not simply find a good starting point and then march along the contour pair forming a 1:1 mapping. The bottom edge of the foam block provided the pair of contour segments used in our running example, above. This image pair is contrived to exhibit significant foreshortening, requiring the point matcher to warp left image contour segments into right contour segments 25% longer. The matcher did quite well along the bottom edge of the block, not quite so well along the top in this fairly stringent test. Remember, no external knowledge of the epipolar condition is available; we rely on information extracted from the image pair itself. Nonetheless, the disparity computed along the top edge is a reasonable approximation to the correct result. The ends of the box, for which there is no discernable foreshortening effect, exhibit constant disparity, consistent with those of the adjoining long segment endpoints. This is most encouraging since we have made no effort in this first version of the matcher to ensure disparity continuity across contour segmentation boundaries. The individual conjugate segment pairs are processed independently; no information is propagated between their respective matching processes.

The system works best for aerial images because they typically have large depth to baseline ratios, short focal lengths, the image-to-image transform is nearly affine over significant areas of the scene, and occlusion is practically nonexistent. In our aerial pair the disparity is nearly constant.
because of the coarse sampling but is not constant at the original digitized resolution of 4096 × 4096. We could use the corresponding contour segments from the finer resolution images to refine the disparity map. Nonetheless, the richness of the scene and the consistency of the disparity map demonstrate the algorithm’s reliability.

V. CONCLUSION

We now have reasonable estimates of the epipolar lines, which can be used as “quasi-epipolar strips” to provide matching constraints in a subsequent refinement step which would eventually lead to a precise orientation. The system is uniquely adaptable. Unlike many rule-based approaches, there is no particular domain assumption to cope with. This is exemplified by the performance on the disparate types of image input presented. Point correspondences are recovered in the presence of almost no a priori information, yet the system could readily be adapted to use any such information which might be provided. Much of the power of the approach stems from its exploitation of high-level perceptual constructs early in the correspondence process. The Gestaltic nature of structural stereopsis is exceptionally powerful and we use this to maximum advantage. Because the DEW algorithm recovers such a large number of conjugate points, it can provide better calibration input (to the robust estimator) than manually selected points, in many cases (see Appendix). Manually selected points may be more precise, but for sparse scenes it may be difficult to acquire enough conjugate pairs. The algorithm naturally invites comparison to the earlier work of Hannah [13], [14]. We note that Hannah’s system has its roots in work done over a decade ago; that system is much farther along its evolutionary path and is more complete than ours. Nevertheless, we expect the structural stereo-DEW approach will eventually prove more robust for a broad range of problem domains.

Hannah’s system is considerably different than ours from a philosophical standpoint, yet several of its more interesting features could be incorporated into our framework. It is an area correlation approach embedded in a sophisticated hierarchical control structure. A technique called back-matching, basically matching right-to-left, checks results and discards blunders. Unconstrained hierarchical matching (camera parameters unknown) provides sufficient conjugate pairs to compute the epipolarity constraint. Constrained hierarchical matching then follows which uses the epipolar condition, now known, to restrict the correspondence searches to one dimension from two. In contrast, our system uses structural stereo to perform the 2-D to 1-D restriction before camera orientation is estimated. Finally, Hannah’s system uses “anchored matching” that imposes a disparity smoothness constraint over neighboring matches in two dimensions, and this is radically different than our disparity smoothness constraint. In fact, the 2-D form of the constraint is difficult to enforce in highly occlusive scenes with large depth discontinuities.

Finally, we would like to point to the significant leverage afforded by the use of structural stereo in concert with dynamic edge warping. For LoG zero crossing contours from an operator of \( \sigma = 3 \), we are readily recovering disparities of up to 160 pixels in the \( x \) direction as well as several pixels in \( y \). In the foam block image we recovered disparities over a 60 pixel range along a single conjugate contour segment pair.

We accomplish this without the benefit of epipolar knowledge and without the burden of coarse-to-fine tracking in a scale space framework. There are no search windows, and no prior disparity estimate is needed across the image pair. We make an “on-line” estimate from the conjugate contour centroids; no external estimate is necessary. The MPG approach would require an initial disparity estimate at each, say, left image point within \( \pm w = \pm 2 \sqrt{2} \sigma \approx \pm 8.5 \) pixels. That is, a 17 pixel search window would be set up along the epipolar line. Tanaka and Kak are using some high level information and, for this case, an 8 or 9 pixel window. We have eliminated the notion of a search window and, to a large extent, the notion of coarse-to-fine tracking altogether. This is a direct consequence of the introduction of high-level perceptual tokens early in the stereopsis process.

APPENDIX

EPIPOLARITY CALCULATION AND DEW-MANUAL COMPARISON

The coordinate system is shown in Fig. 12 with origins at the lens centers, image planes at \( z = 1 \) (\( z' = 1 \)), and unit focal lengths. Object space is thus \( z < 0 \). Let \( P \) be a point in object space represented by \((x, y, z)\) in the left camera coordinates and \((x', y', z')\) in the right. Then:

\[
[x', y', z'] = R_0 [x, y, z] + T_0
\]

where \( R_0 \) is a \( 3 \times 3 \) orthonormal rotation matrix and \( T_0 \) is a \( 3 \times 1 \) translation vector. We refer the reader to [9] for a thorough discussion of the robust estimation of these parameters from a set of point correspondences; theirs is the method employed here.

Calculating the Epipolar Lines

Epipolar lines in the two images are defined by the family of planes passing through the optical centers of the two cameras and intersecting the two image planes. Each epipolar line has a corresponding line in the other image along which its correspondences lie. We now derive equations for such lines from the estimated rotation and translation matrices. We will parameterize the family of epipolar planes, and compute the line of intersection of each plane with each of the two image planes, yielding a parameterized set of correspondences.

Let us first derive the parametric equations for the epipolar lines in the left camera. Using (9) we see that the center of the right camera is along the vector \( F = R_0 T \) with respect to the left camera. Therefore the right camera’s optical center is at \( kF \) where \( k \) is a scaling constant. The
equation of a plane passing through the optical center of the left camera, \([0,0,0]\) and that of the right camera, \(kF = [k_f_1, k_f_2, k_f_3]\), is given by:

\[
m_1x + y = z\left(\frac{f_1}{f_3} + \frac{f_2}{f_3}\right).
\]

(10)

The intersection of this family of planes, parameterized by \(m_1\), with the image plane, \(z = 1\), defines the epipolar lines in the left camera:

\[
y = -m_1x + \left(\frac{f_1}{f_3} + \frac{f_2}{f_3}\right).
\]

The epipolar planes defined by (10), can be found in the right camera coordinate system using the transformation defined by (9). The family of epipolar planes is defined by

\[
\frac{m_1r_{11}^p + r_{12}^p - r_{13}^p c_1}{m_1r_{11}^p + r_{12}^p - r_{13}^p c_1} = \frac{y'}{z'} = \frac{m_2r_{11}^p + r_{12}^p - r_{13}^p c_1}{m_2r_{11}^p + r_{12}^p - r_{13}^p c_1},
\]

(11)

where \(r_{ij}^p = R^p(i,j)\) and \(c_1 = (m_1f_1 / f_3 + f_2 / f_3).\) (Ideally, \(R^p = R^p_0\) if \(R_0\) is truly orthonormal.) Intersecting this family of planes with the image plane, \(z' = 1\), defines the epipolar lines as seen in the right camera. Thus the family of corresponding epipolar lines are defined by the following equations:

\[
y = -m_1x + c_1
\]

\[
y' = -m_2x + c_2
\]

\[
m_2 = \frac{m_1r_{11}^p + r_{12}^p - r_{13}^p c_1}{m_1r_{11}^p + r_{12}^p - r_{13}^p c_1}
\]

\[
c_1 = \left(\frac{f_1}{f_3} + \frac{f_2}{f_3}\right) m_1; \text{ parameter of the family.}
\]

(12)

Relative Orientation from DEW and Human Correspondence

We will now present the results of camera pose estimation comparing point correspondences computed by the DEW algorithm with those identified by hand. Because the calibration algorithm is an iterative reweighted least squares method, it can accommodate blunders in the correspondence data. This is especially important when the errors may not be independent and identically distributed. The nature of the DEW algorithm is such that systematic errors may result. We present the results as estimates of the camera to camera transformation matrices and as estimates of the epipolar lines, plotted back onto the corresponding edge images.

The calibration test image pair is that of the foam block, Fig. 8. These images were taken with a pair of cameras hanging from a common, roughly horizontal crossbar on a floor-to-ceiling gantry. The cameras were set up with axes more or less parallel, but not necessarily orthogonal to the crossbar.

The DEW algorithm identified 794 conjugate points between this pair of images, all around the silhouette of the foam block. The image plane size is 9.6 mm × 12.8 mm and the focal length is 12.5 mm. The estimated rotational and translational matrices are:

\[
\hat{R} = \begin{bmatrix}
0.9892 & 0.0812 & 0.1222 \\
-0.0154 & 0.999 & -0.0043 \\
0.0178 & 0.0977 & 0.9951
\end{bmatrix},
\]

\[
\hat{t} = \begin{bmatrix}
-0.8429 \\
0.0495 \\
0.5359
\end{bmatrix}.
\]

Notice that the rotation matrix (ideally orthonormal) is nearly the identity, which is what we would expect, given the camera arrangement as described above. Without precise control, we have no way of knowing what the exact rotation matrix should be, but this is obviously close. Given the angular relationship (as we observed it) between the optic axes and the gantry crossbar, the translation matrix (ideally a unit vector) is reasonable, as well.

By hand we were able to identify only 14 conjugate pairs with reasonable confidence, which illustrates another strength of the DEW algorithm (which found 794 by virtue of its fine scale examination of the image structure) for this purpose. The camera pose estimation results in this case are:

\[
\hat{R} = \begin{bmatrix}
0.9690 & -0.01075 & 0.2227 \\
0.0336 & 0.9994 & 0.0105 \\
-0.1855 & 0.1080 & 0.9767
\end{bmatrix},
\]

\[
\hat{t} = \begin{bmatrix}
0.8322 \\
0.0130 \\
0.5540
\end{bmatrix}.
\]

One's first inclination is probably to consider the human-selected result as the "ground truth" and consider the discrepancies to be essentially due to erroneous correspondences from the DEW algorithm. This, however, is not the case. The DEW algorithm has, in fact, outperformed the human operators in this experiment. That this is the case can be seen by considering Fig. 13. In the top pair of images, (a) and (b), we have plotted the epipolar lines computed from the DEW relative orientation estimate back onto each of the two edge images. Also plotted on the two images are the conjugate points identified by hand, but these were not available to the DEW algorithm or the robust estimator. Corresponding epipolar lines have common labels. The identification of corresponding points was successful enough to allow reasonable, although not yet highly precise, relative orientation of the camera pair. In particular, notice where epipolar lines 7, 10, and 14 meet the upper right, lower left, and lower right corners of the silhouette, respectively, in each image. Further notice how epipolar line 2 is nearly collinear with the short, upper edge of the silhouette in each case. Perhaps most significant, however, is how well the hand-selected points, which were not used in this calibration, align along corresponding epipolar lines.

In Figs. 13(c) and (d) we plot the epipolar lines obtained from the hand selected conjugate points back onto the edge images in the same manner as above. First of all, notice that the epipolar lines "fan" in the opposite direction, suggesting that the relative pan angle between the two cameras is reversed. This appears in the rotation and translation matrices as sign changes in several of the positions. If one now compares the relationships of corresponding epipolar lines to the underlying edge images, as we did for the DEW-driven results above, we see that the calibration result here is much poorer. In particular, notice the behavior along the left vertical edge of the block and the corners between that edge and the long sides. Also, notice that the calibration result is also poorer for the hand selected points which drove it. Although we selected the hand correspondences very carefully (admittedly, we are not experts in that regard), the results are simply not as good as those obtained from the DEW input. The reason is obvious, the DEW algorithm submits such a large number of points to the calibration routine, that
the set is, in effect, rich; the conjugate points are well scattered throughout the volume of the jointly visible object space. The robustness of the estimator "forgives" a reasonable number of blunders that may be present, especially given the large number of data points to work with. Of course, given a richer image pair, and a skilled operator, hand-selected correspondences should eventually prove superior for calibration in a single step. We believe, however, that a successive refinement mechanism will eventually allow this approach to perform as well as the manual technique, regardless of scene content.

ACKNOWLEDGMENT

We are grateful to the Intergraph Corporation, particularly Dr. Jon Dykstra, for support and helpful discussions. We would also like to thank the reviewers for numerous insightful comments and suggestions that have improved the paper considerably.

REFERENCES


Kim L. Boyer (M'86) received the BSEE (with distinction), MSEE, and Ph.D. degrees, all in electrical engineering, from Purdue University, West Lafayette, IN, in 1976, 1977, and 1986, respectively.

From 1977 through 1981 he was with the Satellite Transmission Laboratory, Bell Telephone Laboratories, Holmdel, NJ. From 1981 to 1983 he was with the Image Processing Department, COMSAT Laboratories, Clarksburg, MD. Since 1986 he has been an Assistant Professor of Electrical Engineering at The Ohio State University, Columbus, where he founded the Signal Analysis and Machine Perception Laboratory (SAMPL). Dr. Boyer's current research interests include stereopsis in weakly constrained environments, optimal feature extraction, inexact structural matching, perceptual organization, and the organization of large structural models bases.

Daniel M. Wuescher received the BEE degree from Cleveland State University in 1987 and the MS in electrical engineering from The Ohio State University, Columbus in 1989.

Since 1989 he has been with the Intergraph Corporation, Huntsville, AL, where his current activities include research and development in aerial image analysis for state-of-the-art computer-aided photogrammetry, cartography, and geographic information systems applications.

Sudeep Sarkar (S'90) received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1988, where he was awarded the general proficiency award as the best graduate in electrical engineering. He also received the National Talent Scholarship from the Government of India. In 1990 he received the MS in electrical engineering from The Ohio State University, Columbus, where he is currently pursuing the Ph.D.

His research interests span machine vision and signal processing.