Translating Optical Flow Into Token Matches and Depth From Looming

Lance R. Williams
Allen R. Hanson

COINS TR 88-68

August 1988

This research is sponsored in part by the Rome Air Development Center (RADC) and the Defense Advanced Research Projects Agency (DARPA) under contract F30602-87-C-0140.

TRANSLATING OPTICAL FLOW INTO TOKEN MATCHES AND DEPTH FROM LOOMING [1]

Lance R. Williams and Allen R. Hanson

Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

Motion in the environment manifests itself in changes of many kinds, not just image plane velocities, yet these are all that an optical flow field makes explicit. We view optical flow as a useful low-level representation from which a symbolic description of change, in the form of token matches, can be computed. The tokens of interest to us are those produced by perceptual organization processes, and are more abstract than edges or interest points. We demonstrate a working system for matching line tokens which uses the optical flow field in a heuristic manner to limit the search for the *minimal bipartite cover* of the set of tokens from each frame. As an example application, we demonstrate a technique for computing distance to environmental surfaces suitable for obstacle recognition by a mobile robot. Accurate knowledge of the camera motion parameters is not required. We describe how motion in depth manifests itself in the projected lengths and areas of environmental surfaces whose extent in depth is small relative to their distance from the camera. Results on two sequences taken by a mobile robot are presented to demonstrate the accuracy of the method.

[1]This research is sponsored in part by the Rome Air Development Center (RADC) and the Defense Advanced Research Projects Agency (DARPA) under contract F30602-87-C-0140.

1 Introduction

It is our position that the inherently local measurement of visual motion provided by optical flow is insufficient to meet the varied requirements of dynamic image understanding. We choose to describe the time varying image by computing correspondence between tokens of arbitrary spatial scale produced by perceptual organization processes. We believe that this will result not only in more accurate measurement of visual motion, but also facilitate the use of motion information in object recognition and scene understanding.

For the purposes of this paper, techniques for measuring visual motion can be roughly categorized as either optical flow methods or token matching methods. This taxonomy is based upon the nature of the output representation.

Specifically, a distinction is drawn between methods whose purpose is the computation of a velocity or displacement field, and methods which compute correspondence in time between tokens that serve as descriptors of spatial structure.

The goal of the optical flow methods is to compute a vector function of the image plane. Depending on the particular method, each vector represents either a velocity or a displacement. The typically pixel-parallel

nature of the computation is dictated largely by the form of the input and output representations, which are arrays of values in registration with the original scene. Optical flow methods have been more successful than token matching methods, and much of this success is due to the fact that these methods implicitly exploit information carried directly in the "shape" of the image intensity surface. This is usually effected through the use of an *intensity constancy constraint* (e.g. See Horn and Schunk [12]). The utility of optical flow methods has been further increased by the ease with which they can be expressed hierarchically, resulting in faster algorithms capable of describing larger motions [2,9,19].

The token matching methods which concern us compute correspondence in time between spatial structures produced by grouping processes. These tokens belong to what Marr has called the *full primal sketch*, [16] and are more abstract than edge segments [17] or interest points [3,18,21]. Tokens map directly to environmental structure, and descriptions of their movement correlate more closely with the motion of physical objects, than does optical flow. Most importantly, token matching allows change through time to be expressed in a wide variety of ways. A token match represents more than a spatial displacement, also explicit are the changing values of any parameters associated with the token. These can include orientation, length,

area, contrast, color, etc. Although this information is explicit in the output of a token matching method, it is difficult or impossible to compute from an optical flow field. These parameters are often solely the products of the particular grouping process responsible for the creation of the token and therefore have no local counterparts. For example, there have been attempts to characterize the rotational component of the optical flow through operators with purely local spatial support such as div and curl [13]. In constrast, knowing the actual value of the orientation of a line token (produced by a grouping process, and possessing arbitrarily large spatial support) as it changes through time results in a more accurate determination of angular velocity. As a further example, in Section 3 we show how precise knowledge of changing lengths and areas of tokens composed of two or more straight line segments allows determination of distance to that structure in the environment.

While spatial structure is better described by a set of tokens than by an array of values, current perceptual organization processes fail to adequately describe the shape of the image intensity surface. Indeed, this information is usually intentionally discarded during the abstraction process. However, even if a sufficiently powerful descriptive language were developed (.e.g. [4,10]), and a token matching approach were formulated, it is difficult to see

how such an approach could rival the efficiency and simplicity of methods which exploit this information implicitly, through an intensity constancy constraint.

The fact that a representation is easy to compute reveals nothing about its utility. Interpretation of an optical flow field, independent of any knowledge of the spatial structure from which it was derived, seems difficult at best. Spatial structure can be characterized locally with an interest operator [18] and interpretation can be restricted to the sparse set of points with a high interest operator score. Alternatively, local structural measures can be incorporated within the optical flow computation itself, allowing a dense flow field to be computed through a *smoothness constraint.* Nagel employs a second order approximation of the intensity variation to determine the direction in which a smoothness constraint is enforced [20]. Anandan analyzes the principle curvatures of the sum-of-squared-difference surface to associate vector confidence measures with each displacement, which in turn, indirectly influence the enforcement of a smoothness constraint [2]. The characterizations of spatial structure in the techniques of Anandan and Nagel are simple and local, as is necessitated by the need to incorporate such characterizations within the formulations of the smoothness constraints.

2 Token Matches From Optical Flow

Recent work in perceptual organization has given us a richer vocabulary with which to describe spatial structure than has been available in the past [15,16,22,27]. While local operators are useful for detecting local structures, more powerful grouping processes are required to recognize structure of arbitrary scale. Since the image structure revealed through grouping corresponds directly with environmental structure [27], perceptual organization provides the best measure of "interest." Recently, Boldt has incorporated these ideas in a working program for grouping long straight line segments [6,24].

With this in mind, a logical course of action might be to enforce a smoothness constraint along the length of a line segment produced by a perceptual organization process such as Boldt's. In fact, Hildreth's smoothness formulation can be enforced along an arbitrary contour, and for the specific case of a line segment in three space undergoing rigid motion, she demonstrates that it yields the physically correct flow [11]. If our goal were to produce a better flow field, this would be a reasonable approach. However, it has already been suggested that knowledge of correspondence in time between the line segments themselves would result in a representation more useful to the interpretation task. We view the optical flow field as a

convenient way to represent the information provided by the intensity constancy constraint for use by the symbolic matching process. It is a useful low-level representation from which a more abstract description of change, in the form of token matches, can be computed.

2.1 Grouping Failure

As tokens become more abstract, they also become more unique. Therefore, it is often assumed that the more abstract a token is, the less ambiguous matching will be. To a certain extent, this is true, but increased abstraction brings with it a new source of ambiguity. Under ideal conditions, we might expect two perceptual organization processes operating independently on two frames of a motion sequence to partition each frame into the same set of tokens. After all, each frame is a slightly different view of a predominantly stable physical world separated only slightly in time. Unfortunately, in practice, due to the discrete sampling of image formation (and other image effects such as shadows, highlights, and texture), the likelihood of two frames being partitioned into the same set of tokens is very small. Since the basic operation employed in perceptual organization is grouping, discrepencies can arise in two ways: 1) failure to relate two tokens that have a single physical cause, or *undergrouping.* and 2) mistakingly relating two

tokens that have separate physical causes, or *overgrouping*. While both undergrouping and overgrouping can be caused by noise, overgrouping errors most often occur when two tokens satisfy the geometric criteria for grouping through pure chance. Unfortunately, this happens more frequently in motion sequences depicting the view of the world from the vantage point of a moving sensor. While the odds of two unrelated tokens accidently satisfying the grouping criteria in any single view are small, the odds of the moving sensor passing through such degenerate views in the course of a motion sequence are much higher. The solution to this problem is beyond the scope of a simple, two-frame matching approach, and it will be discussed in greater detail when suggestions for a multiple frame approach are presented later in the paper.

Even relatively simple undergrouping errors, rule out the possibility of a one-to-one mapping between tokens from successive frames, since the number of tokens in each frame will rarely be the same. Under these circumstances, it seems that the best mapping possible is a mapping that assigns each token at least one match, and optimizes some error function in the process. Such a mapping is called a *minimal bipartite cover*. We first encountered the minimal bipartite cover, in the context of the correspondence problem, as part of Ullman's *minimal mapping theory* [28]. Interestingly,

the motivation Ullman gives for its use is unrelated to the grouping failure argument presented here. We believe that the minimal bipartite cover is simply a more practical goal than a one-to-one mapping when matching abstract tokens prone to grouping errors.

2.2 Frame-to-Frame Token Matches

Ullman's minimal mapping theory, which presents the correspondence problem as an optimization problem in the abstract, is a very general paradigm, and it serves as a useful point of departure for the discussion of the method proposed in this section. In the minimal mapping theory, correspondence is computed between tokens from two frames by finding the minimal bipartite cover of the graph whose nodes are the tokens from each frame and whose arcs reflect potential correspondence. The weight of each arc in the graph is called an *affinity* measure, and is a function of the relative similiarity and spatial separation of the two tokens which the arc links. Ullman justifies his choice of particular affinity values with data from studies of the human visual system. Indeed, the minimal mapping theory is offered as a possible explanation for the manner in which many of the classic Gestalt displays such as Ternus' configuration are interpreted by the human visual system. Because of the explosively large number of possible mappings for matching

problems of even modest size, Ullman simplifies the general matching problem by assuming that the number of candidate matches that each token can claim is equal to some small integer constant. He then shows, that under this assumption, the optimization problem can be solved by a hill climbing process, which leads to a relaxation algorithm. However, no method for choosing initial candidate matches is offered, and the number of iterations required for convergence is unclear.

The approach described in this paper reflects a natural synthesis of the optical flow and token matching paradigms. Because the optical flow field is a vector function of the image plane, it can be used to define a transformation that maps tokens from one frame to their predicted positions in the next frame. The spatial area that must be searched for a potential match is reduced to a small region surrounding the token's predicted position. Furthermore, the merit of a potential match is judged by its proximity to its predicted position, not to its previous position. This simplifies the general matching problem by restricting the number of candidate matches.

2.3 An Implementation Using Line Tokens

In this section, we describe an implementation of the general approach just outlined. A schematic view of the implementation is depicted in Figure 1.

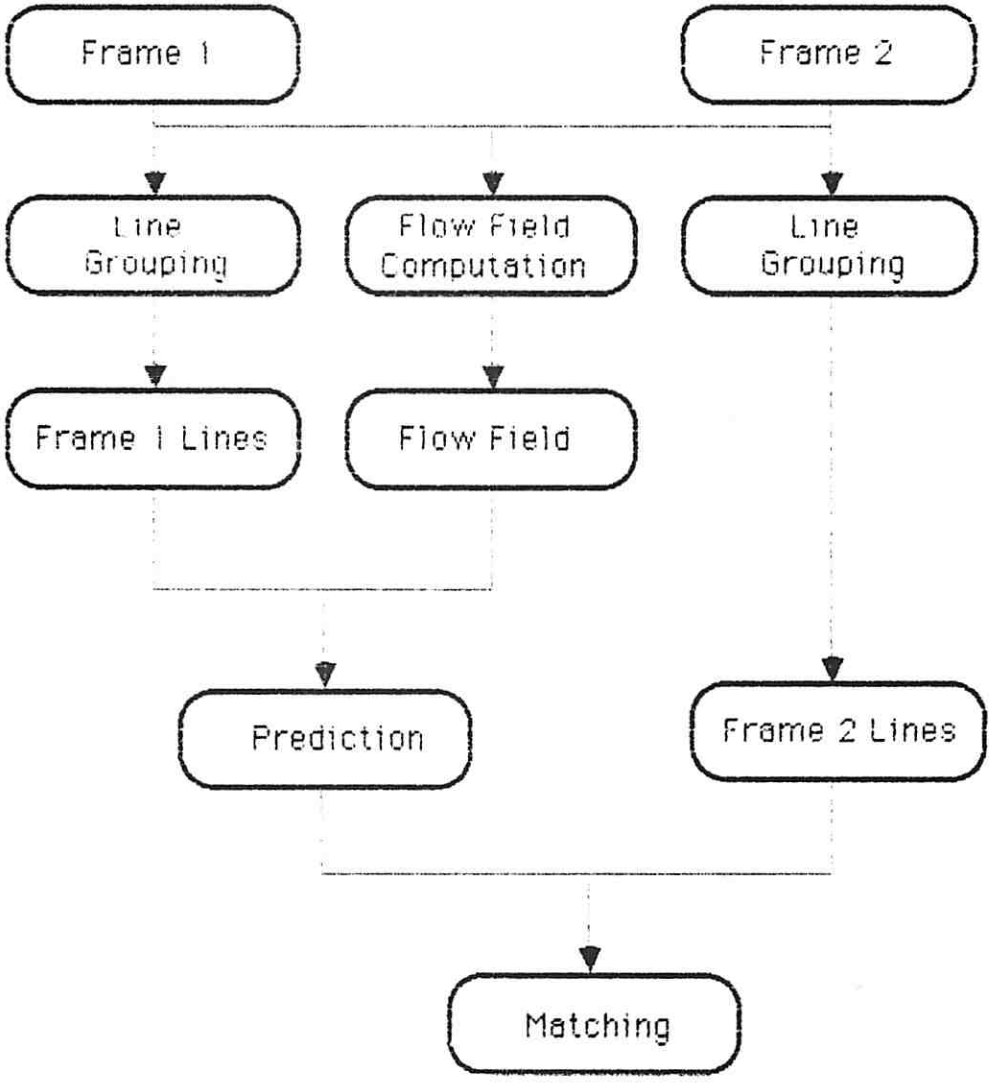


Figure 1. Information flow diagram for an implementation using line tokens.

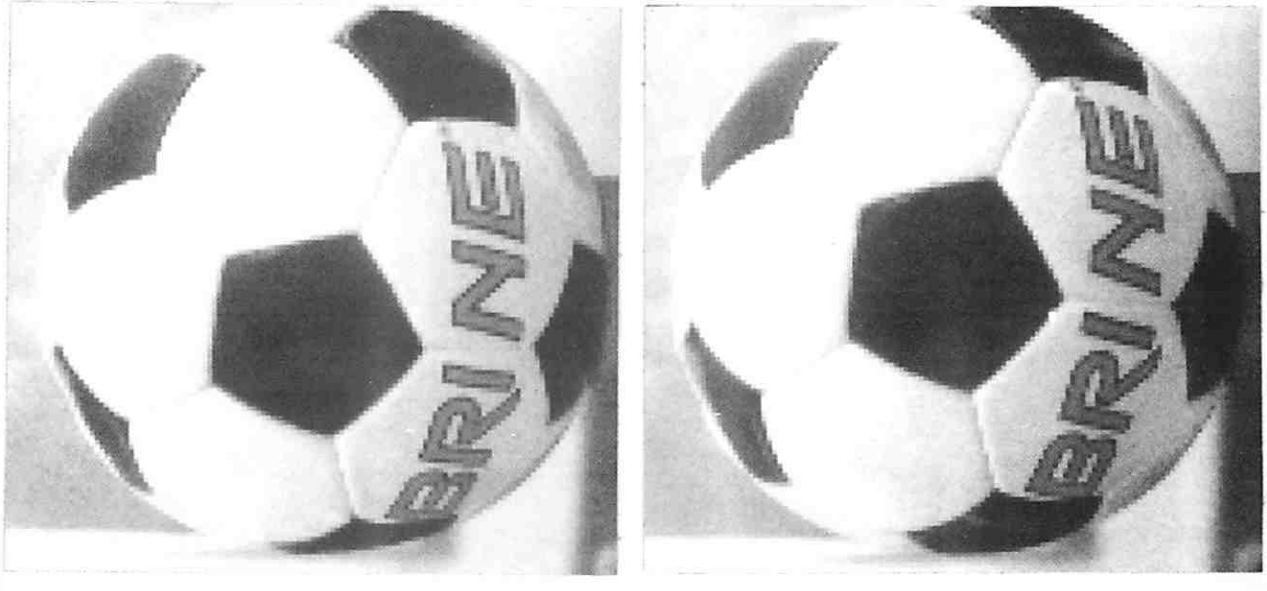


Figure 2. Frame one and two of a sequence depicting a moving soccer ball.

In a preliminary step, a set of line tokens is computed for each frame using Boldt's line grouping algorithm (Figures 2 and 3). Boldt's algorithm begins by extracting an initial set of line segments whose orientation is the normal to the gradient direction along zero crossing contours of the Laplacian operator. These initial line segments form the nodes of a graph whose arcs (*links* in Boldt's terminology) reflect a significant non-accidental geometric relationship between the two line segments they join. Some of the relations used as *linking criteria* are endpoint proximity, orientation difference, lateral distance, overlap and contrast difference. All paths through the *link graph* within the current ***replacement radius*** are examined and the path minimizing the mean-square-error of a straight line fit is replaced by a new line segment. The program is then invoked recursively on the new set of line segments, using a larger replacement radius, resulting in ever smaller sets of increasingly longer lines. A final set of between one and two hundred lines is produced by filtering on length and contrast.

In a second preliminary step, the optical flow field is computed using the method developed by Anandan [2] (Figure 4). Strictly speaking, Anandan's algorithm produces a displacement field, not a flow field. The intensity constancy constraint exists implicitly as a sum-of-squared-difference measure within a Laplacian pyramid. Anandan's algorithm uses knowledge of the

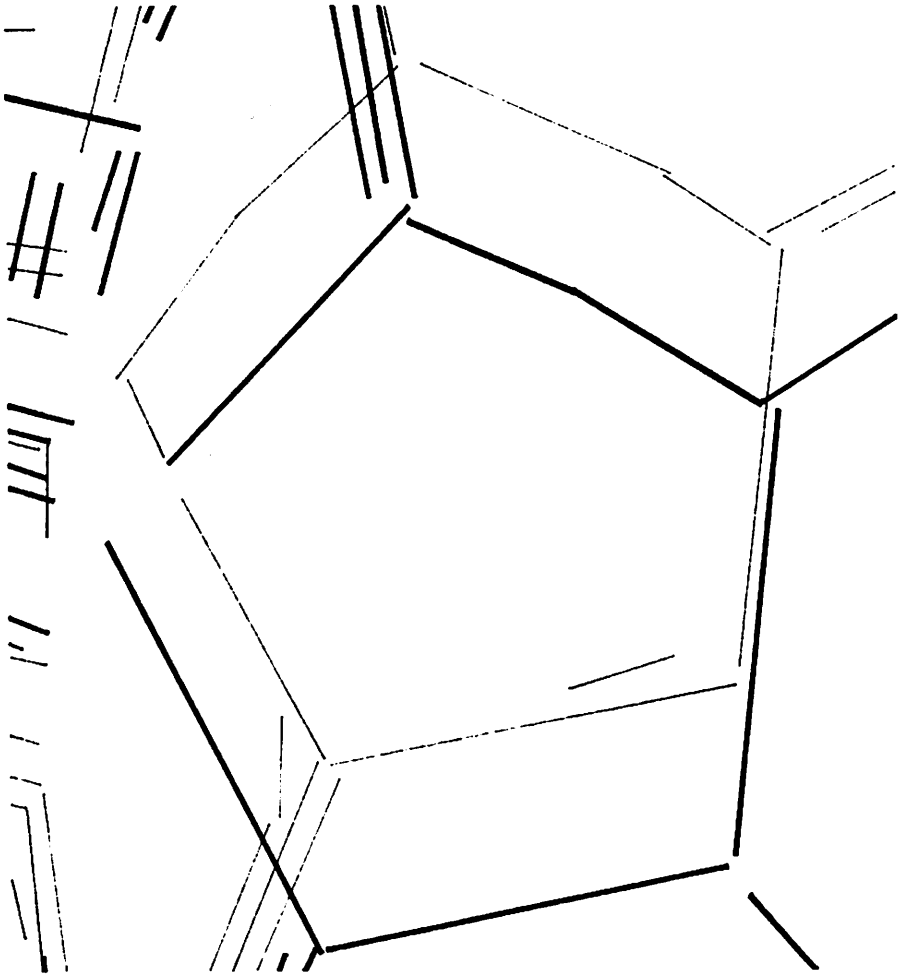


Figure 3. Line tokens computed with Boldt's grouping algorithm. Lines for frame one are displayed thick while lines for frame two are displayed thin.

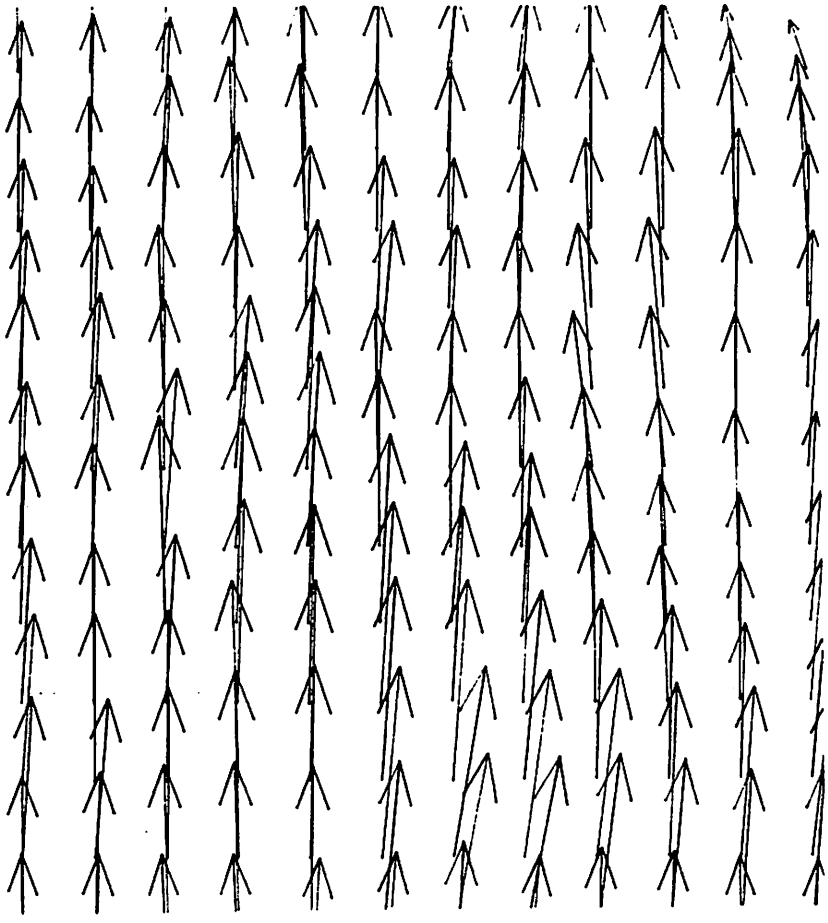


Figure 4. Sample of the displacement vectors computed with Anandan's algorithm.

direction of principle curvature of the sum-of-squared-difference surface to enforce a smoothness constraint at each level of the Laplacian pyramid. These design choices together comprise a working system that appears to consistently yield reliable estimates of image displacements.

The predicted position for each line from the first frame is computed by a least squares fit to the points comprising the image of that line under the transformation defined by the optical flow field. All lines from the second frame passing through a narrow rectangular region surrounding this predicted position are retrieved (Figure 5). The size of the search region is a parameter of the system. Although currently a constant, it could conceivably be coupled to the value of confidence measures associated with the optical flow, such as those computed by Anandan's algorithm. In this way, the window size would be smaller in areas of high confidence and larger in areas of low confidence.

A bipartite graph, henceforward called the *time-link graph*, is constructed. Its arcs connect line segments from the frame one token set, to all candidate matches retrieved from the second frame (Figure 6). The weight of each arc in the time-link graph is a measure of the discrepancy in position between the predicted position of the frame one line segment and the position of the candidate match. Since the line segments' lengths are highly

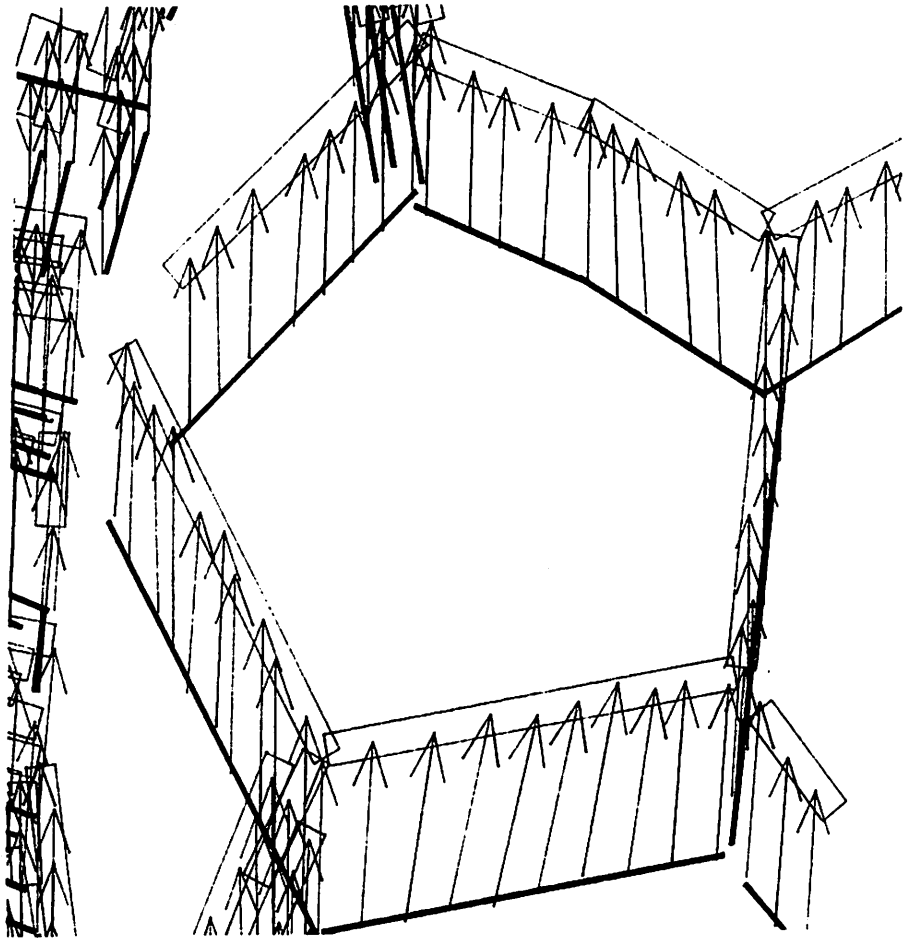


Figure 5. The rectangular search regions computed for each line token by a least squares fit to the tips of the displacement vectors. Lines from frame two that intersect the search region become candidate matches.

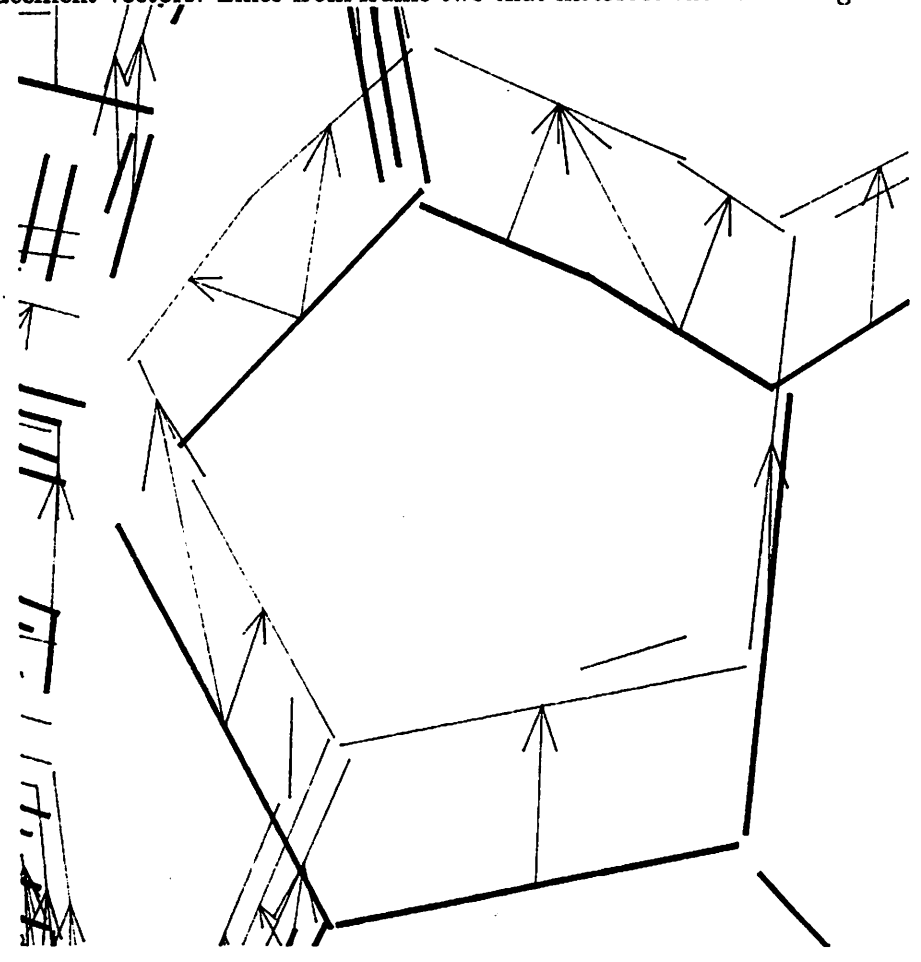


Figure 6. The *time-link graph*, where the arcs connect lines in frame one with their candidate matches.

unstable (because of undergrouping errors) information about length is not incorporated into the positional discrepancy metric. Instead, the measure approximates the average distance between the two segments, independent of their lengths (Figure 7). Ideally, one would use a measure similiar to that employed by Lowe in his model matching system [15], which computes the probability of the juxtaposition of two line segments being due to chance alone, using knowledge of the distribution of background line segments.

By computing the connected-components of the time-link graph, the global matching problem is conveniently divided into smaller, individually tractable pieces which reflect the scope of potential interactions. For each connected-component, the bipartite cover minimizing the positional discrepancy metric is found. This is accomplished through a simple blind search of the sub-graphs of each connected-component. Although Ullman suggests solving the optimization problem through network relaxation, the need for such an approach is eliminated here because of the relatively small size of each connected-component. Indeed, a connected-component often contains only a single arc, in which case the match is uniquely determined. This is directly due to the heuristic use of the optical flow field. The bipartite cover reflects the final correspondences reported by the system (Figure 8).

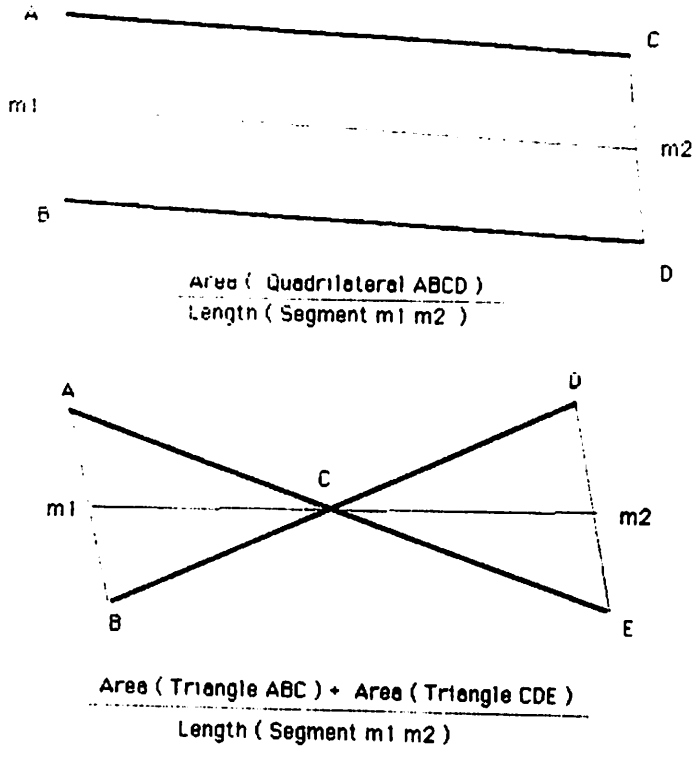


Figure 7. The positional discrepancy metric approximates the average distance between the candidate match and the predicted position of the frame two line.

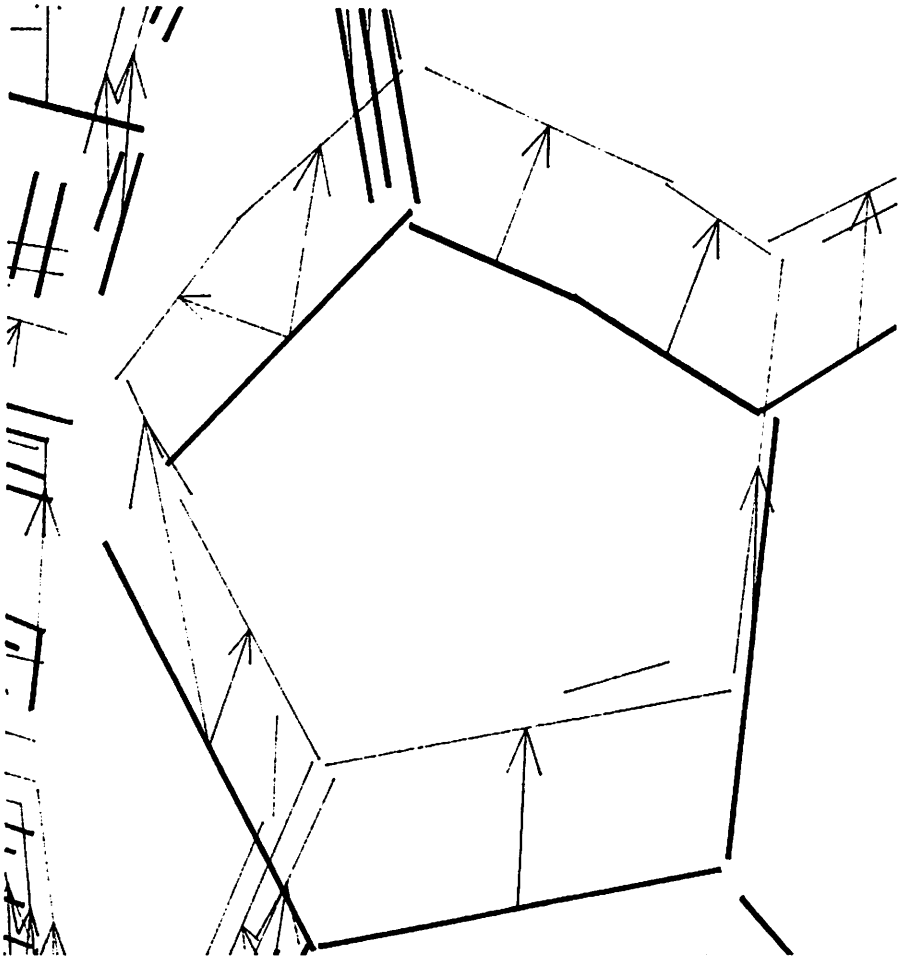


Figure 8. The *minimal bipartite cover* for each connected-component of the time-link graph. In this example, a single arc has been removed.

This system has been used to process more than fifty different two-frame matching problems drawn from six different multi-frame sequences. All the sequences are composed of images of real scenes and contain more than one hundred lines each. No special attention was paid to the magnitude of the displacements between frames, and the system seems reasonably robust to the problems posed by undergrouping errors. Encouraged by the quality of the two-frame results, the system was run repeatedly on successive frames of a multi-frame sequence, creating a *directed acyclic graph*, or *dag*, representing the splitting and merging of line segments over time. The results of one such multi-frame experiment, involving several rotating objects, are shown in Figures 9-12. Results from a second sequence, taken by a camera mounted on a mobile robot panning by a stairway, are shown in Figures 13-16.

Unfortunately, the interpretation of such a representation is non-trivial. For example, one can not tell from local information alone whether a particular split or merge in the dag is due to an undergrouping or overgrouping failure. Although the minimal bipartite cover functions well when the set of line segments in each frame is the same (except for fragmentation due to undergrouping) it performs badly when wholly new line segments appear or dissapear. This can be caused by the initial filtering operation used to

Figure 9. The first frame of a motion sequence containing multiple independently moving objects.

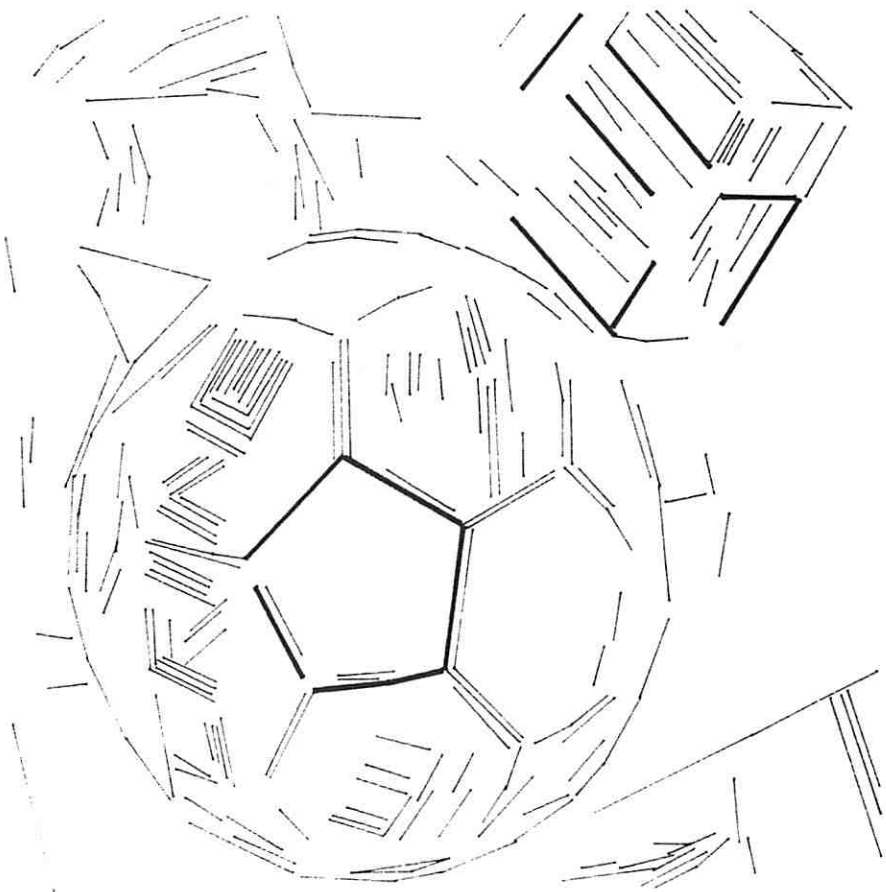


Figure 10. The line tokens computed for the first frame. Line tokens which will be used to illustrate the output of the matching process are displayed thick.

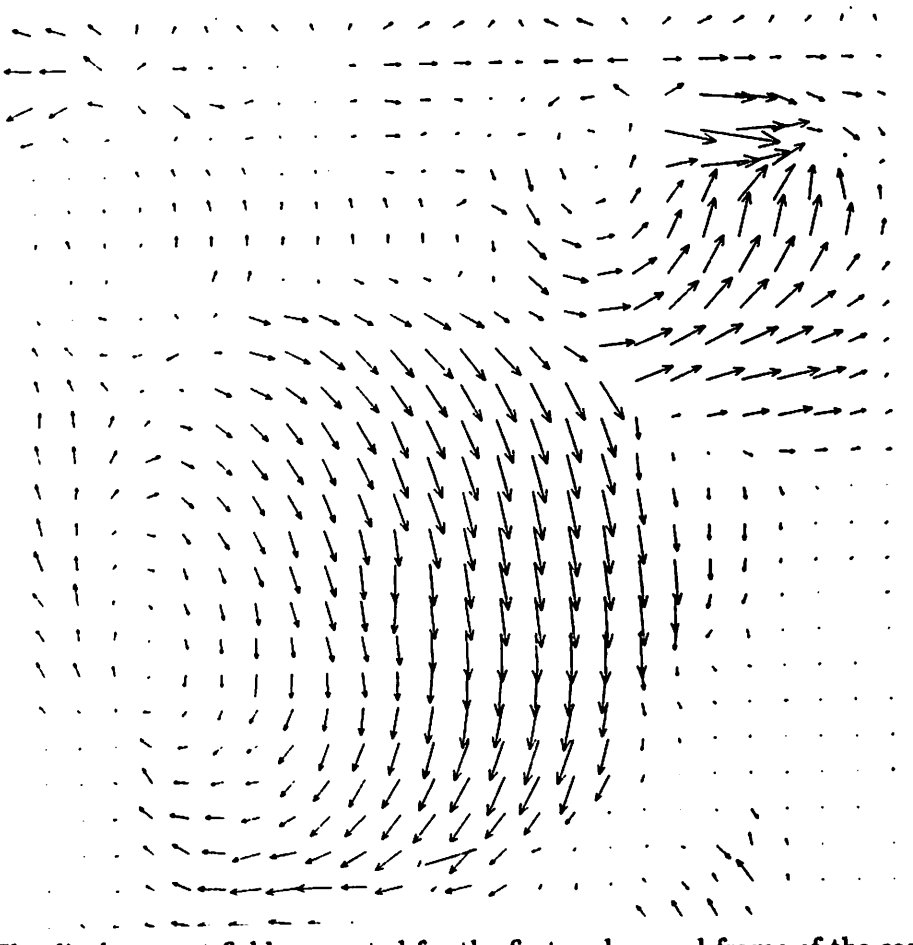
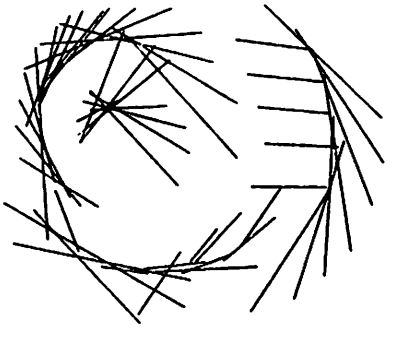


Figure 11. The displacement field computed for the first and second frame of the sequence. Note the rotation of the box and the soccer ball.



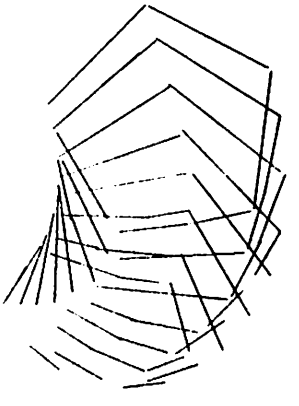


Figure 12. The output of the matching process for selected lines.

Figure 13. The first frame of a motion sequence taken by a mobile robot panning by a stairway.

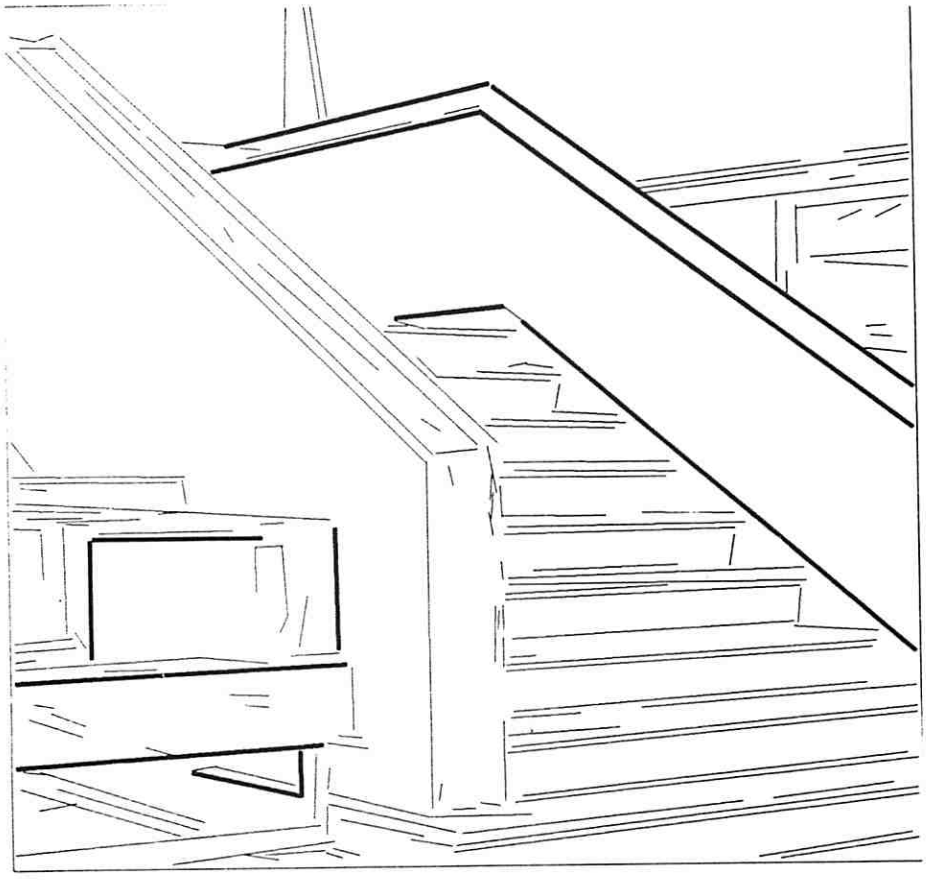


Figure 14. The line tokens computed for the first frame. Line tokens which will be used to illustrate the output of the matching process are displayed thick.

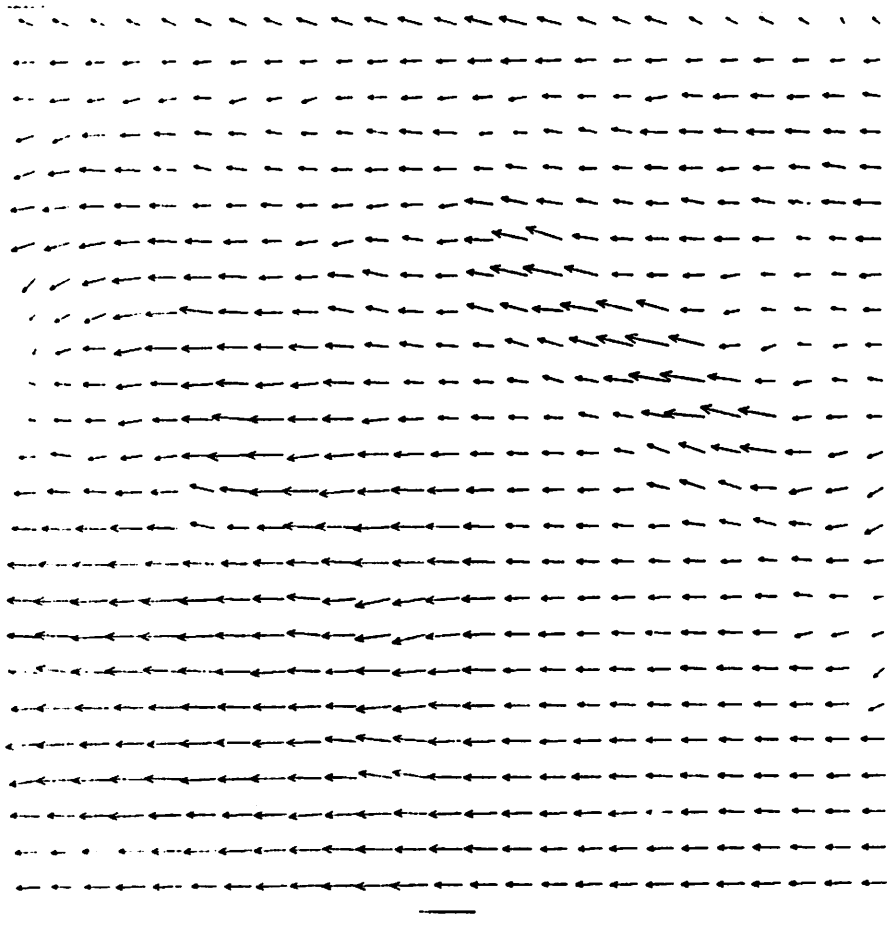


Figure 15. The displacement field computed for the first and second frame of the pan sequence.

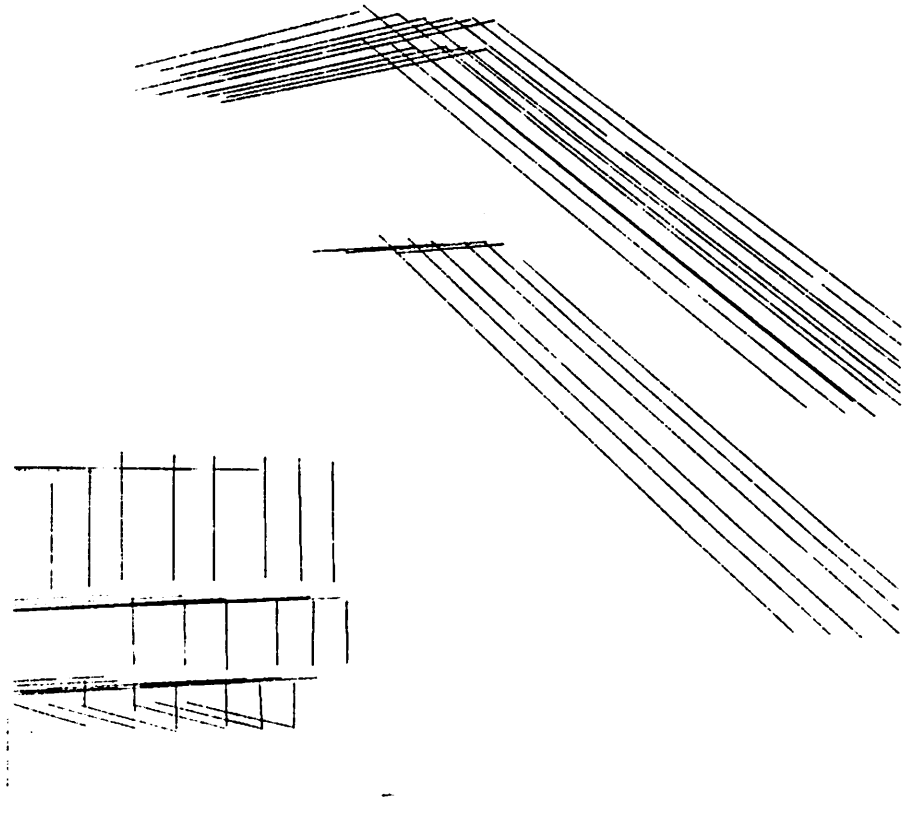


Figure 16. The output of the matching process for selected lines.

reduce the number of line segments in each frame.

2.4 P.O. In Parameter Space

As mentioned before, overgrouping occurs when, through chance, two tokens are juxtaposed in such a way as to satisfy the requirements for grouping. For example, two coplanar line segments, will appear colinear when viewed from any point in the plane in which they both lie (i.e. the *degenerate view plane*). Such a pair of segments is likely to satisfy the grouping requirements in an algorithm such as Boldt's, and will be grouped as a single line segment; See Figures 17 and 18. The probability of a moving sensor passing through the degenerate view plane for some pair of lines is relatively high, especially in a man made environment, due to the plethora of horizontal and vertical lines. However, if we choose to describe each line as a point in the $\rho - \theta$ parameter space, and examine the set of such points through time, we will find two distinct trajectories that intersect during the degenerate view. The appropriate solution seems to be to divide the set of points in parameter space into distinct trajectories corresponding to separate physical entities. This is a perceptual organization problem, and the space to be organized is defined by the parameters of the token. For point tokens, the parameter space happens to be the image plane, but for line

Figure 17. The fifteenth frame of a multiframe sequence depicting a *degenerate view* of line segments on the staircase and doorway. The line segments and the camera are coplanar in three space.

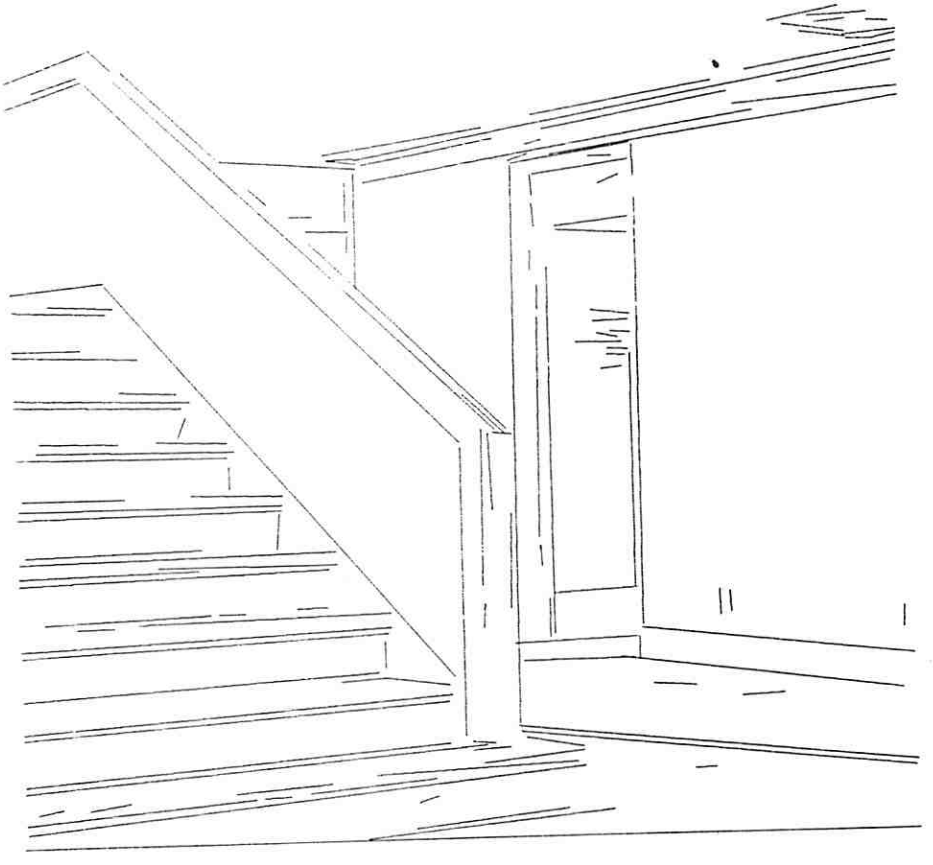


Figure 18. The *overgrouping error* resulting from the accidental satisfaction of the grouping criteria for Boldt's algorithm.

segments, the most suitable parameters might be ρ and θ, which are stable even when undergrouping errors make the determination of the segment's endpoints impossible. Perceptual organization of line token trajectories in $\rho - \theta - t$ space has not yet been fully explored. However, by extending the "perceptual radius" to a larger number of frames, we hope to take advantage of good continuity, which will permit matching in spite of single frame grouping errors.

3 Depth From Looming Structure

The remainder of this paper describes a representative application, in which the line token matches are used to recover depth to environmental surfaces.

Perspective geometry tells us that the optical flow which confronts an organism moving through its environment is purely a function of the motion of the organism and the distance to surfaces in the environment. In principle, precise knowledge of the nature of the motion would allow its effects to be subtracted, and the distance to environmental surfaces to be computed. In this way, the organism establishes a relationship with its environment [8]. Conveniently, the motion of the organism itself, or the *egomotion*, can, in principal, be computed from the optical flow.

The problem of computing the parameters of the egomotion is vastly

simplified when it is known, *a priori*, that the motion is purely translational. For the case of pure translational motion, knowledge of the position of the *focus of expansion*, or *FOE*, provides two of three translation parameters, the third being velocity in the direction of gaze. There have been several attempts to use this assumption for computing distance to environmental surfaces [14,5]. Unfortunately, although this assumption is sound in theory, even a very small deviation from pure translational motion (in the form of rotation) results in significant errors in the determination of the position of the FOE (See [7]). This renders techniques which rely on accurate knowledge of the position of the FOE effectively useless.

Although there has been some success in computing distance to environmental surfaces assuming completely general motion [1,7], depth values computed in this manner are, likewise, only as accurate as the estimates of the egomotion parameters. The goal of computing distance to environmental surfaces without full and accurate knowledge of the egomotion parameters remains attractive.

In this section, we derive equations illustrating how motion in depth manifests itself in the projected lengths and areas of environmental surfaces whose extent in depth is small relative to their distance from the camera. We then present results of an experiment with an image sequence from the

mobile robot domain to demonstrate the potential accuracy of the method.

3.1 The Time Adjacency Relation

Perspective projection can be approximated by a scaled orthographic projection when two conditions are met [23]. First, the depth to the centroid of the environmental structure in question must be large with respect to the focal length of the camera. Second, the total extent in depth of the structure must be small compared to the depth of its centroid. We call an environmental structure satisfying these two requirements a *shallow structure*. We assume that for shallow structures, scaled orthographic projection and perspective projection are equivalent. Assuming that environmental structure, of length L, satisfies the shallow structure requirement and lies at a distance, z, from the image plane, then its projected length, l_0 will be

$$l_0 = \frac{Lf}{z} \tag{1}$$

where f is the focal length of the camera.

If the imaging device is translating into the environment with velocity, $\vec{T}$, then the component of the velocity in the direction of gaze, T_z, is

$$T_z = \vec{T} \cdot \hat{z} = |\vec{T}| \cos\theta \tag{2}$$

where, θ is the angle between the direction of gaze and the FOE. Thus,

knowledge of the position of the FOE is required only to compute the component of $\vec{T}$ in the direction of gaze. Since T_z is proportional to $\cos\theta$, and since $\cos\theta$ is essentially equal to one when the FOE and the direction of gaze are close (which is normally the case), errors of several degrees in the determination of the position of the FOE can be tolerated. After time t, the projected length l_1 will be

$$l_1 = \frac{Lf}{z - T_z t} \tag{3}$$

From this, we see that the ratio, $\frac{l_0}{l_1}$ is

$$\frac{l_0}{l_1} = \frac{z - T_z t}{z} \tag{4}$$

Solving for z, we get

$$z = \frac{T_z t}{1 - \frac{l_0}{l_1}} \tag{5}$$

This is essentially the *time adjacency relation*,

$$\frac{z}{T_z t} = \frac{l_1}{l_1 - l_0} \tag{6}$$

where l_0 and l_1 are not the distances from the FOE of a point at two different times, but are rather the lengths of the projection of some environmental structure. We can thus view the time adjacency relation of [14] as a special

case of the looming structure relation; the FOE and the point in motion define the projection of an imaginary line segment.

We can generalize the looming structure relation for projected lengths to projected areas. Assuming that environmental structure, with area A, satisfies the shallow structure requirement and lies at distance, z, from the image plane, then its projected area, a_0 will be

$$a_0 = \frac{Af^2}{z^2} \tag{7}$$

After time, t the projected area will be

$$a_1 = \frac{Af^2}{(z - T_z t)^2} \tag{8}$$

As with lengths, we compute the ratio, $\frac{a_0}{a_1}$ and see that it is

$$\frac{a_0}{a_1} = \frac{(z - T_z t)^2}{z^2} \tag{9}$$

Solving for z, we get

$$z = \frac{T_z t}{1 - \sqrt{\frac{a_0}{a_1}}} \tag{10}$$

3.2 Experimental Results

The sequences used to demonstrate these ideas were taken with a camera mounted on a mobile robot and have rotational components large enough

to frustrate an algorithm dependent on accurate knowledge of the position of the FOE (Figure 19). The line matching results are satisfactory, although the lengths of the line segments produced by Boldt's algorithm [6] (or any grouping algorithm) are often unreliable. Fortunately, the orientation and lateral placement of the lines is accurate, and we exploited this fact to define *virtual lines* whose length could be accurately measured over the course of the motion sequence. The endpoints of the virtual lines are defined by the intersections of two pairs of line segments. Knowledge of the correspondence through time of the line segments defining the virtual lines, provides information about the changing parameters of the virtual lines themselves. Just as virtual lines can be defined by two pairs of physical lines, *virtual regions* can be defined with three or more pairs (Figure 20). For these experiments, virtual lines and regions satisfying the shallow structure requirement were defined manually, through a graphic user interface. It is our goal to eventually automate this process by exploiting general organizational principles such as endpoint proximity, convexity, symmetry, etc.

Although the virtual lines and regions used in this experiment are either intensity discontinuities or are bounded by them, this is not a requirement. The virtual lines and regions used appear with labels in Figures 21 and 22.

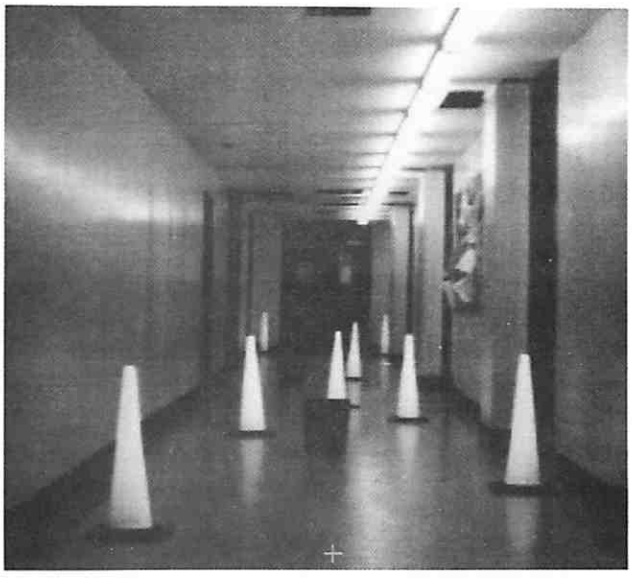


Figure 19. The first frame of a motion sequence taken by a mobile robot moving down a hallway.

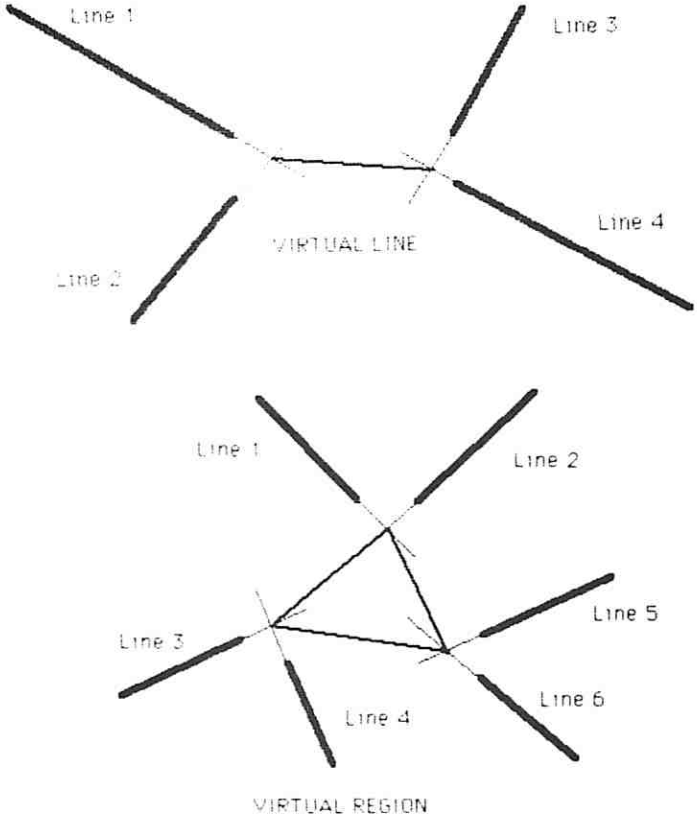


Figure 20. Defining *virtual lines* and *virtual regions* with pairs of line segments.

Table 1.

Virtual Line	Depth (ft.)	Ground Truth (ft.)	% Error	t
Cone 1	19.1	20.0	4.5	1
Cone 2	23.6	25.0	5.6	3
Cone 3	28.3	35.0	19.1	1
Cone 4	42.1	40.0	5.3	7
Can 1	29.0	30.0	3.3	7
Wall 1	27.7	27.1	2.2	2
Wall 2	48.8	48.7	0.2	7
Doorway	88.8	87.1	2.0	7

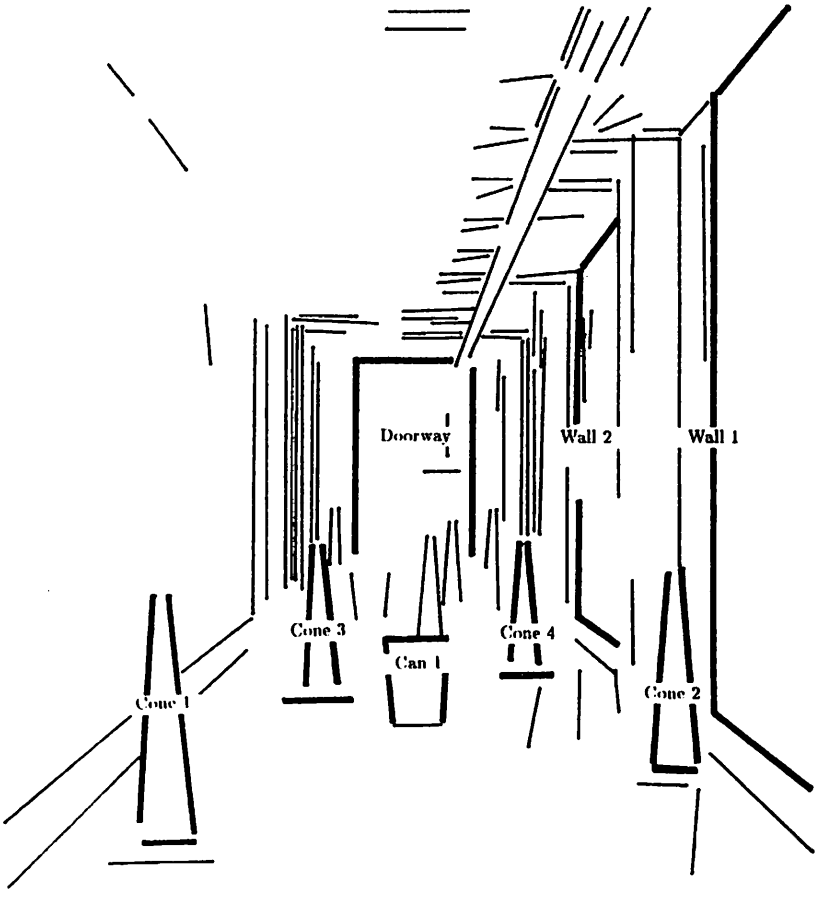


Figure 21. The line segments used to define *virtual lines*.

Table 2.

Virtual Region	Depth (ft.)	Ground Truth (ft.)	% Error	t
Cone 1	20.1	20.0	0.5	1
Cone 2	25.8	25.0	3.2	3
Cone 3	35.5	35.0	1.4	1
Cone 4	40.0	40.0	0.0	7

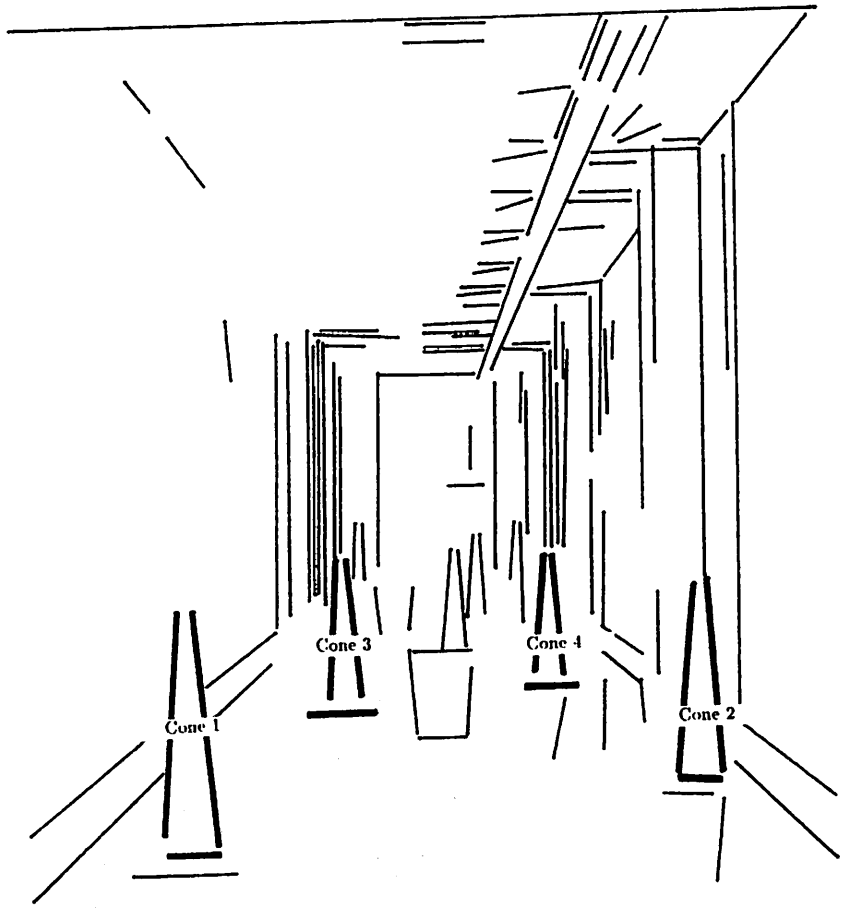


Figure 22. The line segments used to define *virtual regions*.

To increase accuracy, each depth value was computed over the largest interval that all line segments defining the virtual line or region were tracked, that is to say, until one line exited the image or failed to have an acceptable match.

Knowledge of the position of the FOE improves the accuracy of depth estimates but is not critical. For this experiment, the position of the FOE was estimated by hand, although algorithms exist which are at least as accurate [14]. The robot moved a distance of 1.95 feet between frames. Knowing the position of the FOE allowed us to estimate T_z, the component of the robot's motion in the direction of gaze, as 1.91 feet. This results in less than a 3% increase in accuracy over simply assuming that the FOE and the direction of gaze are identical. The depths to the virtual lines are shown in Table 1, along with the ground truths, percent errors and the number of frames contributing to the depth estimate. Table 2 displays the same information for the virtual regions.

The looming method was tested on a second sequence taken by the mobile robot (Figure 23). The aggregate structures used are slightly different than those employed in the hallway sequence. Figure 24 shows several four point configurations which form virtual regions whose changing area can be measured. The depths are displayed in Table 3. Figure 25 shows sev-

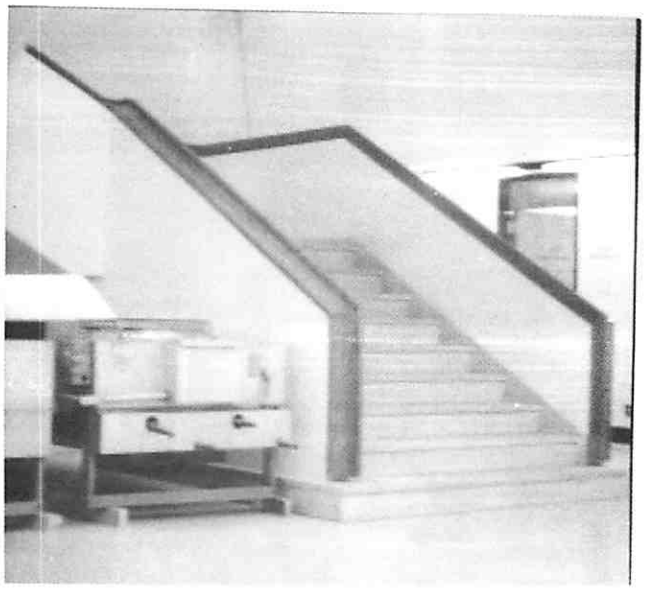


Figure 23. The first frame of a motion sequence taken by a mobile robot moving towards a stairway.

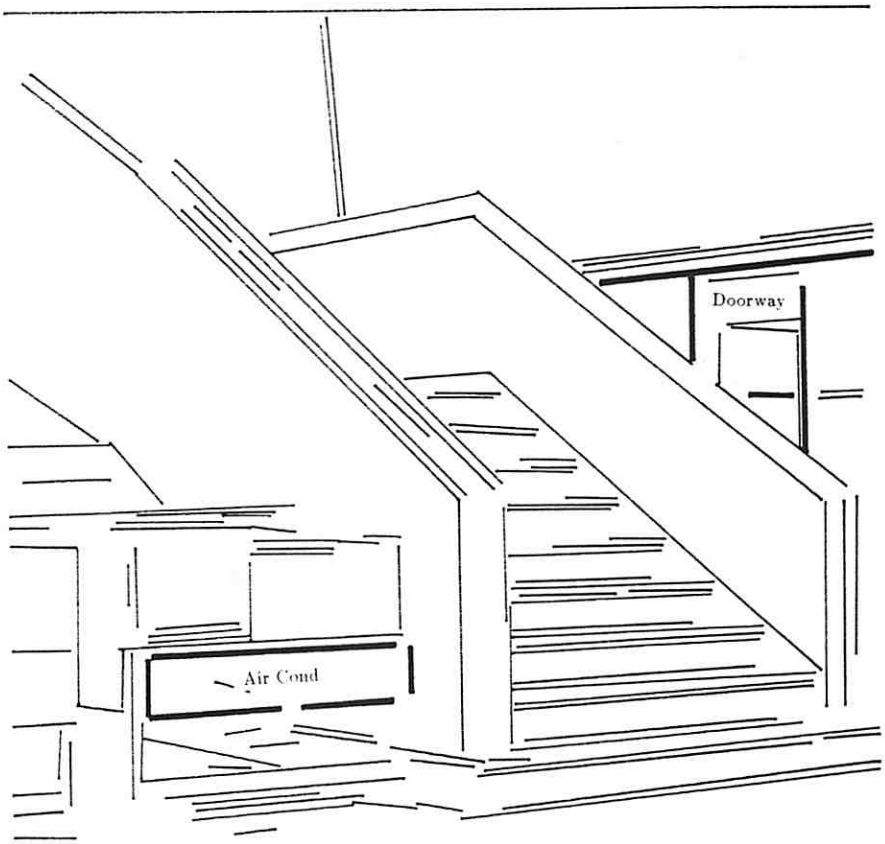


Figure 24. Four point configurations used to define virtual regions, whose changing area can be measured.

eral pairs of parallel lines, or *bars*, whose changing width can be measured. The depths are displayed in Table 4. Since many of the structures are not parallel to the image plane, the computed depths and ground truths are approximate, and interpretation of the results is more subjective than in the case of the hallway sequence.

4 Conclusion

An optical flow field is a vector function of the image plane. It is a very simple characterization of the changing intensity function that results when a dynamic scene is imaged. Compared to token matching, it is a relatively well developed paradigm and several different algorithms exist for computing it. The first part of this paper explores the possibility of translating optical flow into token matches, creating a more abstract representation of motion based on a directed acyclic graph. The nodes of this graph are tokens corresponding to spatial structure and the arcs reflect correspondence between frames. In addition to the spatial displacement of the token, this representation makes the changing values of the token's parameters explicit. The approach is demonstrated by a working implementation which uses line tokens. Finally, it is proposed that the best path to pursue in future work is perceptual organization in the parameter space of the token. Hopefully,

Table 3.

Virtual Region	Depth (ft.)	Ground Truth (ft.)	% Error	t
Air Cond	23	23	0	1
Doorway	42	43	2	6

Table 4.

Bar	Depth (ft.)	Ground Truth (ft.)	% Error	t
Air Cond 1	22	23	5	1
Air Cond 2	21	23	10	1
Stair 1	27	25	8	6
Stair 2	29	29	0	6
Doorway 1	44	43	2	6
Doorway 2	47	43	10	6

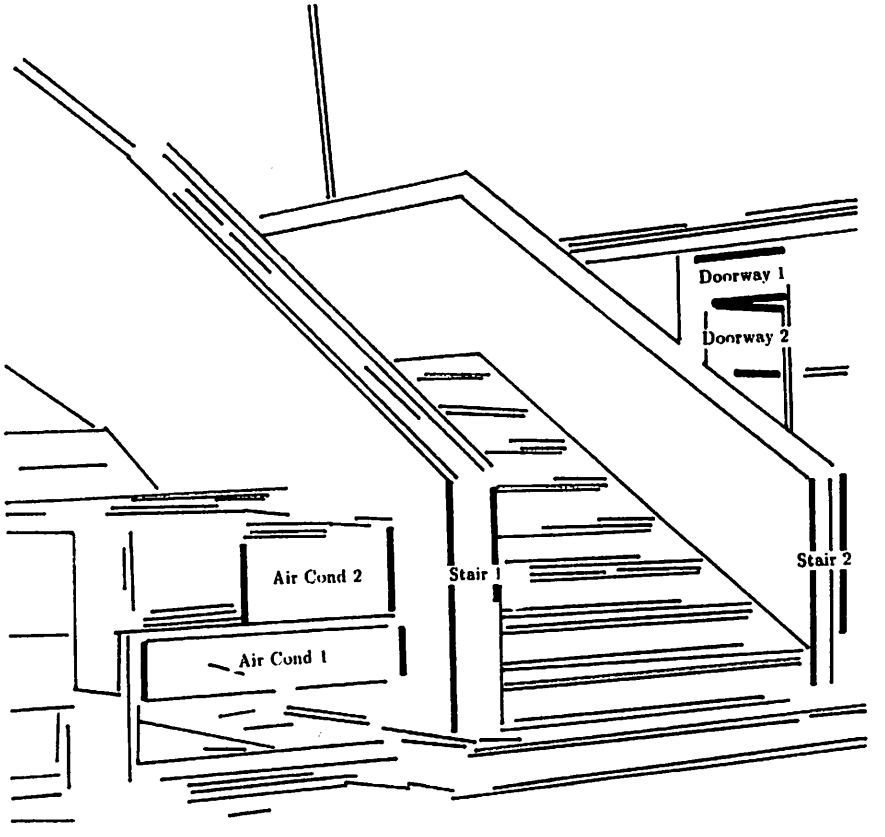


Figure 25. Pairs of parallel lines, or *bars*, whose changing width can be measured.

this will provide increased reliability in the face of single frame grouping errors.

Finally, as an example application, we used the token matches generated with the line matching algorithm to demonstrate that depth to environmental surfaces can often be computed from a motion sequence without first completely determining the egomotion parameters. Depth information manifests itself not only in image plane velocities, but also in the changing lengths and areas of structural descriptors. A simple formulation for the special case of environmental structure whose extent in depth is small compared to its distance from the camera has been derived. The potential accuracy and utility of the "looming" method has been demonstrated in experiments with image sequences from the mobile robot domain.

ACKNOWLEDGMENTS

We wish to thank Michael Boldt, P. Anandan, Harpreet Sawhney, Rich Weiss, Ed Riseman and Rakesh Kumar for many valuable discussions and for providing the tools that made this work possible.

References

[1] Adiv, G., Interpreting Optical Flow, Ph.D. Dissertation, COINS Dept., University of Massachusetts, Amherst, Mass., September 1985.

[2] Anandan, P., Measuring Visual Motion from Image Sequences, Ph.D. Dissertation, COINS Dept., University of Massachusetts, Amherst, Mass, March 1987.

[3] Barnard, S.T. and Thompson, W.B., Disparity Analysis of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-2, Number 4, July 1980, pp. 333-340.

[4] Besl, P.J. and Jain, R.C., Segmentation Through Symbolic Surface Descriptions, *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, Miami Beach, Fla., 1986, pp. 77-85.

[5] Bharwhani, S., Riseman, E. and Hanson, A., Refinement of Environmental Depth Maps Over Multiple Frames, *Proceedings of the IEEE Workshop on Motion*, 1986.

[6] Boldt, M. and Weiss, R., Token-Based Extraction of Straight Lines, COINS Technical Report 87-104, University of Massachusetts, Amherst, Mass, October 1987.

[7] Dutta, R., *et al*, Issues in Extracting Depth from Approximate Translational Motion, *Proceedings of DARPA Image Understanding Workshop*, Cambridge, Mass., 1988.

[8] Gibson, J., *The Senses Considered as Perceptual Systems*, Houghton-Mifflin, Boston, Mass., 1966.

[9] Glazer, F., Hierarchical Motion Detection, Ph.D. Dissertation, COINS Dept., University of Massachusetts, Amherst, Mass, January 1987.

[10] Haralick, R., Laffey, T. and Watson, L., The Topographic Primal Sketch, *The International Journal of Robotics Research*, Spring 1983.

[11] Hildreth, E.C., The Measurement of Visual Motion, Ph.D. Dissertation, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, Mass., 1983.

[12] Horn, B.K.P., and Schunck B.G., Determining Optical Flow, *Artificial Intelligence*, Vol. 17, pp. 185-203.

[13] Koenderink, J.J., and Van Doorn, A.J., Invariant Properties of the Motion Parallax Field Due to the Movement of Rigid Bodies Relative to an Observer, *Optica Acta*, Vol. 22, No. 9, pp. 773-791., 1975.

[14] Lawton, D., Processing Dynamic Image Sequences from a Moving Sensor, Ph.D. Dissertation, COINS Dept., University of Massachusetts, Amherst, Mass., February 1984.

[15] Lowe, D., *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, Boston, Mass., 1985.

[16] Marr, D., *Vision*, Freeman Press, San Francisco, Cal., 1982.

[17] Marr, D. and Ullman, S., Directional Selectivity and its Use in Early Visual Processing, *Proceedings of the Royal Society of London*, B211, pp. 151-180, 1981.

[18] Moravec, H.P., Towards Automatic Visual Obstacle Avoidance, *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 1977.

[19] Nagel, H.H. and Enkelmann W., An Investigation of Smoothness Constraints for Estimation of Displacement Vector Fields from Image Sequences, *IEEE Transactions on PAMI*, Vol. PAMI-8, pp. 565-593, 1986.

[20] Nagel, H.H., On the Estimation of Dense Displacement Vector Fields from Image Sequences, *Proc. of ACM Motion Workshop*, Toronto, Canada, pp. 59-65, 1983.

[21] Prager, J.M. and Arbib, M.A., Computing the Optic Flow: The MATCH Algorithm and Prediction, *Computer Vision, Graphics and Image Processing*, Number 24, pp. 271-304, 1983.

[22] Stevens, K. and Brookes, A., Detecting Structure by Symbolic Constructions on Tokens, *Computer Vision, Graphics, and Image Processing 37*, pp. 238-260, 1987.

[23] Thompson, D. and Mundy, J., Three Dimensional Model Matching From an Unconstrained Viewpoint, *Proceedings of the IEEE Conference on Robotics and Automation*, Raleigh, N.C., 1987.

[24] Weiss, R. and Boldt, M., Geometric Grouping Applied to Straight Lines, *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, Miami Beach, Fla., 1986, pp. 489-495.

[25] Williams, L.R. and Hanson, A.R., Depth From Looming Structure, *Proceedings of the DARPA Image Understanding Workshop*, Cambridge, Mass., 1988.

[26] Williams, L.R. and Hanson, A.R., Translating Optical Flow into Token Matches, *Proceedings of DARPA Image Understanding Workshop*, Cambridge, Mass., 1988.

[27] Witkin, A., and Tenenbaum, J., What is Perceptual Organization For?, *Proceedings of AAAI '83*, pp. 1023-1026.

[28] Ullman, S., *The Interpretation of Visual Motion*, MIT Press, Cambridge, Mass., 1979.