A COMPUTATIONAL FRAMEWORK AND AN ALGORITHM FOR THE MEASUREMENT OF VISUAL MOTION

P. Anandan

COINS Technical Report 87-73

August 1987

A Computational Framework and an Algorithm for the Measurement of Visual Motion

P. Anandan *
Computer Science Department
Yale University
New Haven, CT 06520

August 12, 1987

Abstract

The robust measurement of visual motion from digitized image sequences has been an important but difficult problem in computer vision. This paper describes a hierarchical computational framework for the determination of dense displacement fields from a pair of images, and an algorithm consistent with that framework. Our framework is based on the separation of the image intensity information as well as the process of measuring motion according to scale. The large scale intensity information is first used to obtain rough estimates of image motion, which are then refined by using intensity information at smaller scales. The estimates are in the form of displacement (or velocity) vectors for pixels and are accompanied by a direction-dependent confidence measure. A smoothness constraint is employed to propagate the measurements with high confidence to their neighboring areas where the confidences are low. At all levels, the computations are pixel-parallel, uniform across the image, and based on information from a small neighborhood of a pixel.

For our algorithm, the local displacement vectors are determined by minimizing the sum-of-squared differences (SSD) of intensities, the confidence measures are derived from the shape of the SSD surface, and the smoothness constraint is cast in the form of energy minimization. Results of applying our algorithm to pairs of real images are included. In addition to our own matching algorithm, we

*The research described in this paper was conducted at the Computer Vision Laboratory at University of Massachusetts, Amherst and was supported by DARPA under grant N00014-82-K-0464

also show that two different hierarchical gradient-based algorithms are consistent with our framework. Moreover, wee point out that the gradient-based techniques (which compute image velocities) can be regarded as mathematical limits of our matching technique. Finally, we discuss some open issues and unsolved problems, especially concerning the processing of discontinuities in image motion.

1 Introduction

1.1 Background

Motion is an important and fundamental source of visual information. It is well known that the pattern of image motion contains information useful for the determination of the 3-dimensional structure of the environment and the relative motion between the camera and the objects in the scene. However, the accurate measurement of image motion from a sequence of real images has proven to be difficult.

While there seems to be widespread agreement that the measurement of motion should be based on primitive image-events (such as intensity fluctuations, points, lines, etc.), and that it should be an early visual process, there seems to be less agreement on its exact definition. In computer vision, the primary emphasis has been on the determination of instantaneous image velocities [17,21,26,34], and the displacements of points between successive frames [8,20], although a few techniques have attempted to track lines and curves [25,43]. Recent developments in psychophysics, however, have focussed on "spatio-temporal energy models" [1,40,41,24] which equate the measurement of motion with the measurement of spatio-temporal energy.

Although many of these techniques are interesting because they are based on solid theoretical foundations, they have usually not been successful in practice. Often, the reason for the difficulty seems to be due to a lack of recognition of the fact that in discrete video sequences, the inter-frame image displacements are often considerably larger than one pixel. In addition, the scene may contain multiple independently moving objects, which means that (i) image motion may not be globally coherent, and (ii) there will be discontinuities in the image velocity (or displacement) field.

In this paper, we describe a hierarchical framework for the computation of dense displacement fields from a pair of images. We also describe a matching algorithm consistent with our framework, and demonstrate its performance when applied to real images. A major contribution of our work is the synthesis of an orientation-selective confidence mea-

sure and a well-defined smoothness constraint within the hierarchical approach, thereby leading to a robust algorithm for measuring large inter-frame displacements. In addition, we have also adapted the "overlapped-pyramid" projection strategy [12] to improve our results, especially around locations of discontinuities in image motion.

Historically, the gradient-based and the matching techniques have been seen as completely unrelated (or even somewhat opposed) to each other. However, we show that the computational framework described here is sufficiently general that hierarchical versions of gradient-based algorithms (which are used for computing image velocities) are also consistent with it. In particular, the recent algorithms of Enkelmann [17] and Glazer [21], which are respectively the hierarchical versions of the second-order technique of Nagel [34] and the first-order technique of Horn and Schunck [26] can be shown to contain all the components of our framework, although some of them in an implicit form.

1.2 Framework overview

The key idea underlying our framework is the separation of computations according to scale. This idea is based on the following observation: usually, the large scale (or low spatial-frequency) intensity variations can provide imprecise measurements over a large range of magnitudes of motion, while the small-scale (or high spatial-frequency) variations can provide more accurate measurements over a smaller range. This leads to three components of our framework: **spatial-frequency decomposition**, which is the method of separating the intensity variations according to scale, a local, parallel **match-criterion** within each scale, and a **control-strategy**, which is a method for controlling the measurement processes at the different scales and combining their results.

Although the scale-based separation of computation provides a useful principle for processing scenes containing large displacements, there will always be situations when an image area lacks sufficient local information for displacement computation at a particular scale. Also, since the image-displacement is a vector quantity, its reliability can vary according to direction. Therefore, another essential component of our framework is a direction-dependent **confidence measure**. The presence of unreliable displacements also means that in order to obtain a dense displacement field, it may be necessary to propagate the reliable displacements to their less reliable neighbors. This leads to the last essential component of our framework: a **smoothness constraint**, which specifies the criterion for the propagation of reliable displacements.

A visual illustration of this framework is provided in Figure 1. The five major components mentioned in the above description are discussed in detail in section 2.

1.3 Paper Organization

The detailed description of our framework for computing dense displacement fields from a pair of images is contained in section 2. First, the goals of the displacement computation are stated. This statement consists of specifying the nature of the input, the requirements on output, and the computational constraints for this process. This is followed by the detailed descriptions of the five components of the framework.

Section 3 describes the specific choices made for the matching algorithm. Of these, the primary areas of our original contribution are the formulation of the confidence measure, the exact method of using the coarse-to-fine strategy, and the formulation of the smoothness constraint. Hence, the description will emphasize these aspects of our algorithm. Section 4 describes a set of experimental results obtained by applying this algorithm to two pairs of real images. While the success of the algorithm is immediately obvious upon visual inspection of these results, our discussion of these results focuses on the failures of this algorithm, so as to indicate the limitations that are inherent to our framework, and to some extent to any low-level approach for the measurement of motion.

Section 5 briefly outlines two hierarchical gradient-based approaches due to Glazer and due to Enkelmann, and explains how these algorithms are consistent with our framework. In addition, section 5 also describes the mathematical relationship between the gradient-based techniques and the matching techniques. In section 6, we briefly discuss the issues involved in processing discontinuities in image motion, while section 7 contains a summary of the main ideas covered in the paper.

2 The Computational Framework

2.1 The computational goals

The goals of the process of computing image displacements are determined by three major factors: the nature of its input, the requirements on the output, and constraints on computational efficiency. The input is a pair of digitized frames belonging to a discrete image sequence. The image displacements may be due to a general 3-dimensional motion of the camera or the independent motion of objects in the scene. The output should be a dense

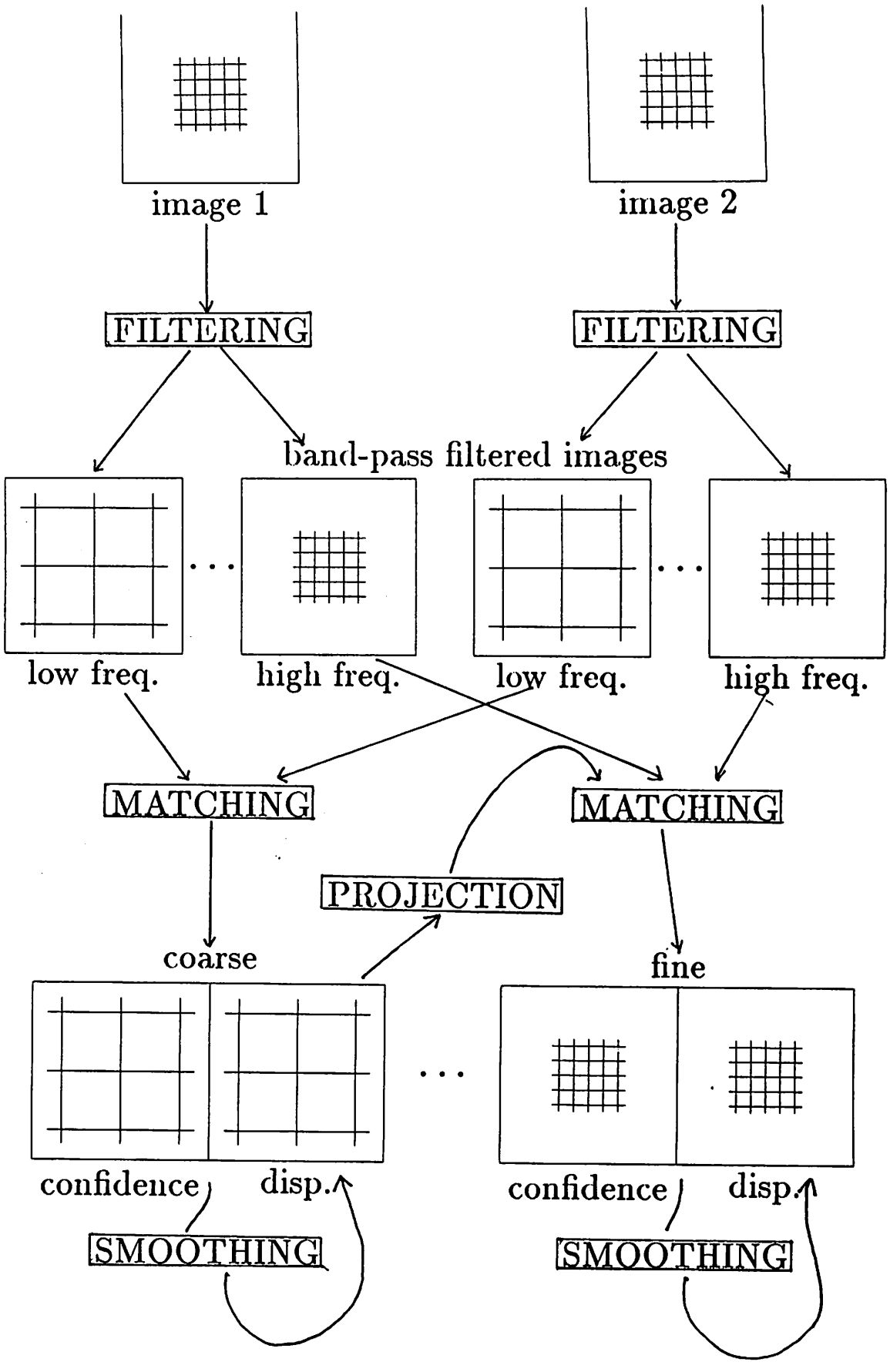

Figure 1: The hierarchical computational framework

field of displacement vectors with associated confidence measures. All the computations must be pixel-parallel and use local image information. Our motivations for choosing this set of goals are explained below.

The input – In typical video sequences, the inter-frame displacements are usually considerably larger than a pixel. Usually, due to the presence of independently moving objects, a single set of 3-D motion parameters will not be consistent with the entire image. It may be assumed, however, that the objects in the environment are composed of continuous and opaque surfaces and that they undergo rigid or nearly-rigid motions. These assumptions mean that that (i) within the image-area covered by a single surface, the displacement field varies smoothly, and (ii) the image-motion can be described as "locally translational", i.e., within a small area of the image, the displacement field can be approximated by a translational flow field. Thus, image sequences containing rotational motion can be processed, as long as the magnitude of rotation between frames is not large. Finally, we also assume that discontinuities in image motion occur at the boundaries of surfaces and objects.

It should be noted that the assumptions of opaqueness and near-rigid motions are essential for the successful performance of any algorithm based on our current computational framework. While it is clear that there will be situations when these assumptions would be violated, such situations are not usual. Besides, in such cases, a "bottom-up" approach may be inadequate for the determination of displacement fields.

The output – The requirement that the output should be a dense displacement field with a confidence measure is derived from the conclusions of the various studies concerning the problem of extracting structure from motion [2,3,18,42]. These studies indicate that a large number of image displacements are necessary for the accurate determination of the structure of the environment. If the scene contains independently moving objects, there may be no *a priori* knowledge of the image-locations of such objects. Therefore, the density of the displacement vector field should be uniformly high across the image, with a confidence measure indicating the reliability of each vector.

Computational considerations – The considerations of computational efficiency and ease of implementation suggest that the following three properties are desirable for all computations: *parallelism*, *uniformity*, and *locality*.

Parallelism simply means that it should be possible to perform all computations simultaneously at all locations on the image-plane. This is required in order to reduce the computational cost. Uniformity implies that the process should be similar at all locations. This is required because of the lack of *a priori* knowledge about image motion. In particular, it should be possible to describe any differences between the computations at different locations in terms of a few simple parameters. Locality means that the computations at any point on the image should be based on information local to that point. This is important in order to reduce the communication cost between processors, as well as to deal with the fact that the displacement field may not be globally coherent.

2.2 Spatial-frequency decomposition

As noted briefly at the beginning of this paper, the key idea underlying the proposed computational strategy is the separation of computation on the basis of scale. Intuitively, it is clear that while small scale intensity structures can be used to measure displacements over a short range, they may have many duplicate matches over a large range. This leads to ambiguities in the computation of the displacements. Therefore, in order to process large displacements, large scale intensity information must be used. However, a single displacement computed on the basis of a large scale intensity structure will be some form of the average of the displacements over the area covered by that structure; hence, its accuracy will be low. Such a "smoothed" displacement field will also vary slowly over the image plane; hence, it can be sampled at a lower rate without loss of information.

These observations lead to the following principle: *large-scale image structures can be used to measure displacements over a large range with low accuracy and at a low sampling density, while small-scale image structures can be used to measure displacements over a short range with higher accuracy and at a higher sampling density.* An obvious way to enforce this principle is to decompose the image into its spatial-frequency components. Such a decomposition and the subsequent processing can be achieved by using a set of *spatial-frequency channels.*

Since the lower-frequency information can be sampled at a lower rate without any significant loss of information, the spatial-frequency decomposition process is usually accompanied by a corresponding reduction of resolution [13,46]. Such an approach leads to a pyramid representation of the spatial-frequency channels and fits naturally into a pyramid [27,37] or a processing-cone [23] architecture. However, since the final choice of a representation scheme depends on the type of hardware which is used, a pyramid representation is

not an essential part of the framework.

2.3 The match-criterion

As noted earlier, the *match-criterion* is a method for determining the displacements within each channel. Since the displacement measured is small with respect to the scale of the intensity variations within a channel, a gradient-based approach can be used (see [17,21]). Alternatively, a correlation-matching approach [6,14,20] or a symbolic matching approach based on primitive tokens [22,29] can also be used. The separation of matching according to scale implies that the match-criterion should have a *scaling* property, i.e., the measurement processes within different channels should be scaled versions of each other. Note that such a scaling property is directly provided in a pyramid representation.

2.4 The control-strategy

The control-strategy determines how the measurement processes at different scales are controlled and how their results are combined. In our framework, the control strategy is based on a *spectral continuity* principle[1] [22,31], which can be described as follows: For images of opaque surfaces, it can be assumed that the displacement estimates at corresponding image locations in the different channels must be similar because they are due to relative motion between the camera and the same environmental area. This means that at any image location, a displacement computed from a high-frequency channel must be consistent with the estimates from the low-frequency channel at the corresponding image location.

A simple way of enforcing the spectral continuity principle is by a "coarse-to-fine" control strategy. In this strategy, the processing proceeds from the low- to the high-frequency channels. The displacement estimate for a pixel in a low-frequency channel determines the center of the search area for the correspoding pixels in the next higher frequency channel. The scale-invariance property of the measurement process suggests that the radius of the search areas between two adjacent channels should be proportional to the scale factor in order to ensure scale invariance of the computations. Once again, note that such scaling is automatically achieved in the pyramid representation.

[1]Note that this principle is usually violated wherever there is a discontinuity in image motion; i.e., in particular at surface and object boundaries. This issue is discussed in greater detail in section 6.

2.5 The confidence measure

In general, there will be many areas of the image with insufficient information at a particular scale for the local determination of displacements. Therefore, a confidence measure should be computed along with each match at each scale to indicate whether or not to accept that match for further processing, or the degree to which the match can be trusted.

Since the image displacement is a vector quantity, it is possible that different directional components of the displacements may be locally computable with different degrees of reliability. For instance, it is intuitively clear that in a homogeneous area of the image no component of the displacement can be reliably estimated. At a point along a line (or an edge), the component perpendicular to the line can be reliably computed, while the component parallel to the line may be ambiguous. Finally, at a point of high curvature along an image-contour it may be possible to completely and reliably determine the displacement vector based on local information. These observations suggest that the confidence measure should be directionally selective, i.e., that it should associate different confidences with the different directional components of the displacement vector. In addition, while an area may be homogeneous at one scale, it may have information useful for reliable matching at a different scale. Hence, the confidence measures should be separately computed within each spatial-frequency channel.

2.6 Smoothness constraint

The computation of a dense displacement field will necessitate "filling in" areas with unreliable displacements based on the reliable displacements in their neighborhood. Such a filling in process can be based on the assumption that the displacement field varies smoothly over the image area covered by a single surface.

The most common use of the smoothness assumption can be found in gradient-based techniques for determining image-velocities. In these techniques, a smoothness constraint is formulated as a variational problem involving the minimization of an error associated with a velocity field. In our framework, we suggest the use of a similar smoothness constraint within each channel. After the displacements and the associated confidences are computed within each channel, the displacement field should be smoothed before it is projected to the next higher frequency channel. During the smoothing process, the confidence measures should be used to retain the reliable displacements while allowing the less reliable estimates to change.

The smoothness assumption is violated at locations of discontinuities in the image motion. Such discontinuities arise at surface boundaries of a single object, or at object boundaries due to independent movement of two different objects. Any scheme that uses the smoothness assumption should also include mechanisms for detecting such violationbs and processing them in an appropriate manner. We consider some approaches for the detection of discontinuities in section 6 of this paper, although it should be noted that, in general, this remains a major unsolved problem.

3 The Hierarchical Matching Algorithm

In this section, we describe a matching algorithm which is consistent with the computational framework. There are five major components to the algorithm, corresponding to the five components of the framework which were described above. Of these, we will outline only briefly the spatial frequency decomposition, the match criterion, and the control strategy, since they are similar to well-known techniques in computer vision. The bulk of this section will focus on the confidence measure and the smoothness constraint, since their exact forms are rather new, and therefore require more detailed discussion. For a detailed description of the entire algorithm the reader is referred to [7].

Any algorithm consistent with a computational framework will also be influenced by the architecture of the machine for which that algorithm is designed. As noted earlier, since pyramid representations and computations are naturally suited for our framework, our primary choice is a pyramid architecture [23,37]. However, it should be noted that we have also developed versions of the algorithm suitable for implementation on a simple mesh-connected-computer [7,45].

For descriptive purposes, we have adopted the convention that the levels of the pyramid are numbered $l = 0, 1, \ldots$, where at any level l, the size of the processor array is $2^l \times 2^l$; level 0 is called the "top" of the pyramid, while level L containing the input image is considered its "bottom".

3.1 Spatial frequency decomposition

A suitable method for the spatial frequency decomposition is provided by the *Laplacian pyramid* transform proposed by Burt[2] [13]. This algorithm consists of two stages: the first stage involves the construction of a Gaussian low-pass filter pyramid from the input image,

[2]See [21] for a discussion of Burt's algorithm and other related algorithms for spatial frequency decomposition.

while the second stage involves computing the difference between the images at adjacent levels of the low-pass pyramid to obtain the set of band-pass filtered images.

The finest level L of the Gaussian pyramid contains the input image. The image at any level $l = L-1,\ldots,0$ is formed by applying a 5×5 Gaussian convolution operation to the image at level $l+1$ and subsampling the filtered image. The Laplacian pyramid image at level l is created by taking the pixel-wise difference between the Gaussain pyramid image at level l and the expanded version of the Gaussian pyramid image at level $l-1$. For our algorithm, one such Laplacian pyramid is constructed from each input image.

3.2 Match criterion

The match criterion chosen was the minimization of a type of correlation measure. Correlation based matching is a traditional and well known approach in computer vision and image analysis. Apart from the fact that such a measure captures the intuitive notion of the similarity (or difference) between two intensity structures, it is also simple to compute and is suitable for parallel and uniform computations. There are several types of correlation measures, e.g., direct correlation, mean normalized correlation, variance normalized correlation, and sum of squared differences (SSD). The definitions and the descriptions of these different measures can be found in [36,4]. An empirical study by Burt, Yen, and Xu [11] comparing these different correlation measures indicated that when used on band-pass filtered images, the computationally simpler measures such as the direct correlation measure and SSD perform nearly as well as the more complex measures. The SSD measure is always positive, a fact which proved particularly useful when we considered the normalization of the confidence measure (described in section 3.4). Hence, our matching algorithm was based on the minimization of SSD.

The exact details of the matching process are as follows: Within each processing level, each pixel in the first image is assigned a a set of candidate matches according to the control-strategy described in secion 3.3. The SSD measure for each candidate pixel is determined by computing the Gaussian weighted sum of the squared differences between the intensity values of corresponding pixels in 5×5 windows centered around the source and the candidate pixel. The best match is selected as the candidate with the minimum SSD.

Although correlation matching can fail dramatically when used in arbitrary situations, it appears to perform reasonably well when it is used within the spatial frequency channels, particularly if the candidate matches are selected according to a hierarchical control

strategy such as ours. Also, an important parameter is the window size. In a simple single-level correlation algorithm, the window size should increase with the width of the search area; otherwise, the likelihood of duplicate matches will increase. However, the use of band-pass filtering, pyramid representation, and the coarse-to-fine control strategy together allow us to use a single fixed window size (in terms of number of pixels) for all points in the images at all levels of the Laplacian pyramid. Our choice of 5×5 window was based on a combination of theoretical and empirical reasons, which are described in detail in [7].

3.3 Control strategy

The control-strategy used in our algorithm is a coarse-to-fine sequential processing of the images in the Laplacian pyramid. In particular, we used an *overlapped pyramid* projection scheme, which is outlined below; a more detailed description can be found in [7].

If the input images are at level L of the pyramid and the maximum displacement of any pixel along either coordinate direction is less than δ, the processing begins at level $L - \log(\delta)$ and proceeds via sequential projection to image level L. At the coarsest level, no displacement exceeds a pixel. Hence, the search area is the 3×3 pixel area centered around the corresponding pixel location in the second image. At all other levels, an initial set of displacements for a pixel are obtained by projecting the displacements from the adjacent coarser level. The search area is the 3×3 area surrounding this projection.

The traditional approach (e.g., see [20]) used for the projection of displacements is based on the quad-tree type of connectivity between adjacent levels of the pyramid. The displacement at a pixel at level $l - 1$ is projected to the four pixels below at level l. However, in this scheme, if the displacement computed for a coarse-level pixel is incorrect, the search areas of none of its descendents at any of the subsequent levels will contain its correct match. Hence, a single match error made at a coarse level causes a large block of pixels at the image resolution to have incorrect displacements.

To overcome some of these problems due to the quad-tree connectivity scheme, we have used an an alternate scheme for projection of displacements, called the overlapped pyramid projection scheme. This scheme was used by Burt, Hong, and Rosenfeld [12] in connection with the problem of image-segmentation. The displacement of a pixel at a coarse-level l is transmitted to all the pixels in a 4×4 area at the next finer level $l + 1$. Thus, each pixel at level $l + 1$ obtains information from four pixels at level l, and can be regarded as

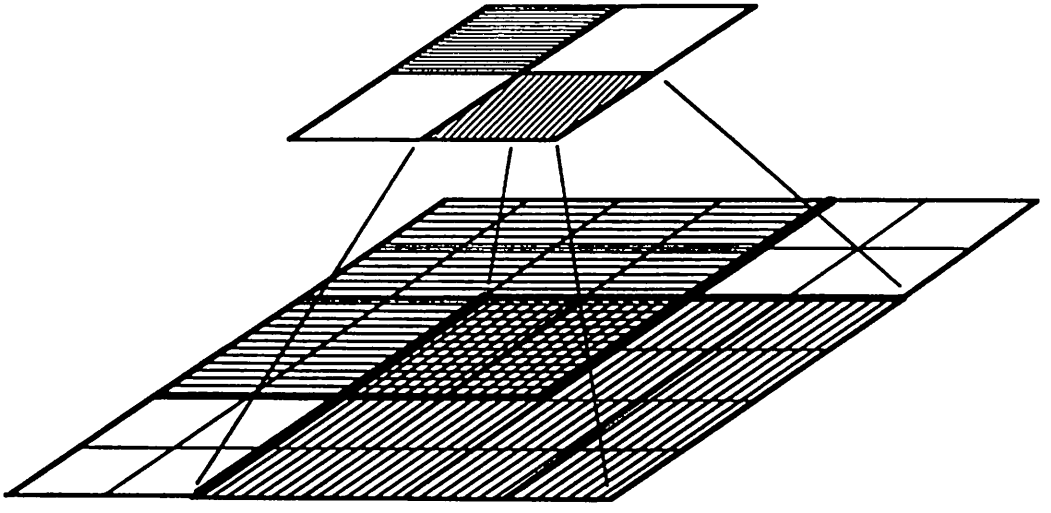

Figure 2: **The Overlapped pyramid projection scheme**

having four potential parents; i.e., the parent of which it is a direct descendent and the three other parents whose projections are adjacent to it (see Figure 2). The displacement of each of the four parents is considered a possible initial estimate for the search at level l; often, however, two or more of these estimates will be identical. The search area consists of the union of the 3×3 areas centered around each distinct coarse-level estimate, and the SSD measure is minimized over all the pixels in this expanded search area.

Although the use of the overlapped pyramid implies that the search area size may increase by up to a factor of 4, the resulting scheme is considerably more reliable. This is because, although a particular coarse-level pixel may be assigned an incorrect displacement, if the displacements of any of its neighbors are correct, the search area for its descendents at the finer-level will still contain their correct matches. A demonstration of the improvement in the results achieved by the use of this scheme can be found in [5].

3.4 Confidence measure

The confidence measure used in our algorithm is based on the variation of the SSD values over the set of candidate matches. Intuitively it is clear that if the value of the SSD measure for different candidate matches are equal then those candidates are equally "good" matches. Thus, if the variation of the SSD measure along a particular line in the search area around the best match pixel is small then all the pixels along that line are equally good matches; i.e., the component of the displacement along the direction of that line cannot be uniquely determined. Also, if the SSD measure corresponding to the best match is large, then it is likely that even the best match is not a reliable match (the implications are discussed later in this section).

Our approach for computing a confidence measure is based on the two intuitions mentioned above. In order to verify that the variations of the SSD measure reflects these properties, we conducted an empirical study involving a pair of synthetic images. This section first describes the results of the empirical study; this is followed by a formal definition of the confidence measure and a discussion its behaviour in various typical situations encountered in real images.

The behavior of the SSD surface

The *SSD surface* is defined over the space of displacements, and its height is the SSD value corresponding to each displacement. For our empirical study, we created a pair of synthetic images by digitally "cutting and pasting" pieces from two real images photographed in the robotics laboratory at the University of Massachusetts. We selected a number of specific points corresponding to typical image structures such as corners, edges, homogeneous areas, and occluded areas, and studied the behavior of the SSD surface computed based on the finest level images of the Laplacian pyramids created from the input images.

The details of our study are too long for this paper, and can be found in [7]. Figures 3, 4, 5 and 6 are examples of SSD surfaces. Figure 3 shows the SSD surface at an intensity corner, Figure 4 shows the surface at an edge, Figure 5 shows the surface at a point in an area of homogeneous intensity, and Figure 6 shows the shape of the surface at an occluded corner point. The surfaces are shown inverted in order to enhance visibility; the viewing position of the surface is noted in each figure and the maximum and minimum elevations of the surface are indicated. In each figure, we have also marked the point of minimum SSD value with the symbol "O", and the point corresponding to the correct displacement

with an "X".

From the figures shown it is evident that at the corner point the SSD surface shows a unique peak (in actuality a minimum), and that the location of the minimum corresponds to the correct displacement. In the case of the edge-like point, a ridge like structure is seen, and the location of the minimum along the ridge seems ambiguous; this suggests that the component of the displacement vector in the direction parallel to the ridge is uncertain. At the homogeneous area the SSD measure shows very little variation, thus suggesting that the selection of a unique match is impossible. Finally, the shape of the SSD surface at the occluded corner point is erratic, indicating that the match may be entirely unreliable.

The types of shapes of the SSD surface illustrated in the accompanying figures, and described above, were observed over many images. In particular, we noted that the curvature of the SSD surface along any direction seemed to indicate the reliability of the component of the displacement vector along that direction. The formal definition of the confidence measure given below is based on these observations.

The definition of the confidence measure

The curvature of a surface at a point along any arbitrary direction can be determined if the two *principal curvatures* at that point and the directions of the associated *principal axes* are known [15]. The principal axes are defined as the directions along which the curvature of the surface is either a maximum or a minimum, and the principal curvatures are the curvatures of the surface along those directions. We denote the maximum principal axis by the unit vector $\hat{e}_{max}$ and the associated curvature as C_{max}, and the minimum principal axis by the unit vector $\hat{e}_{min}$ and the associated curvature as C_{min}.

Our confidence measure consists of two magnitudes (called "confidences") c_{max} and c_{min}, and two directions (or unit vectors) $\hat{e}_{max}$, and $\hat{e}_{min}$. As above, the unit vectors denote the principal axes of the SSD surface, while

$$c_{max} = \frac{C_{max}}{k_1 + k_2 S_{min} + k_3 C_{max}}$$

and

$$c_{min} = \frac{C_{min}}{k_1 + k_2 S_{min} + k_3 C_{min}}$$

where k_1, k_2, and k_3 are parameters that are normalization parameters, and S_{min} is the SSD value corresponding to the best match.

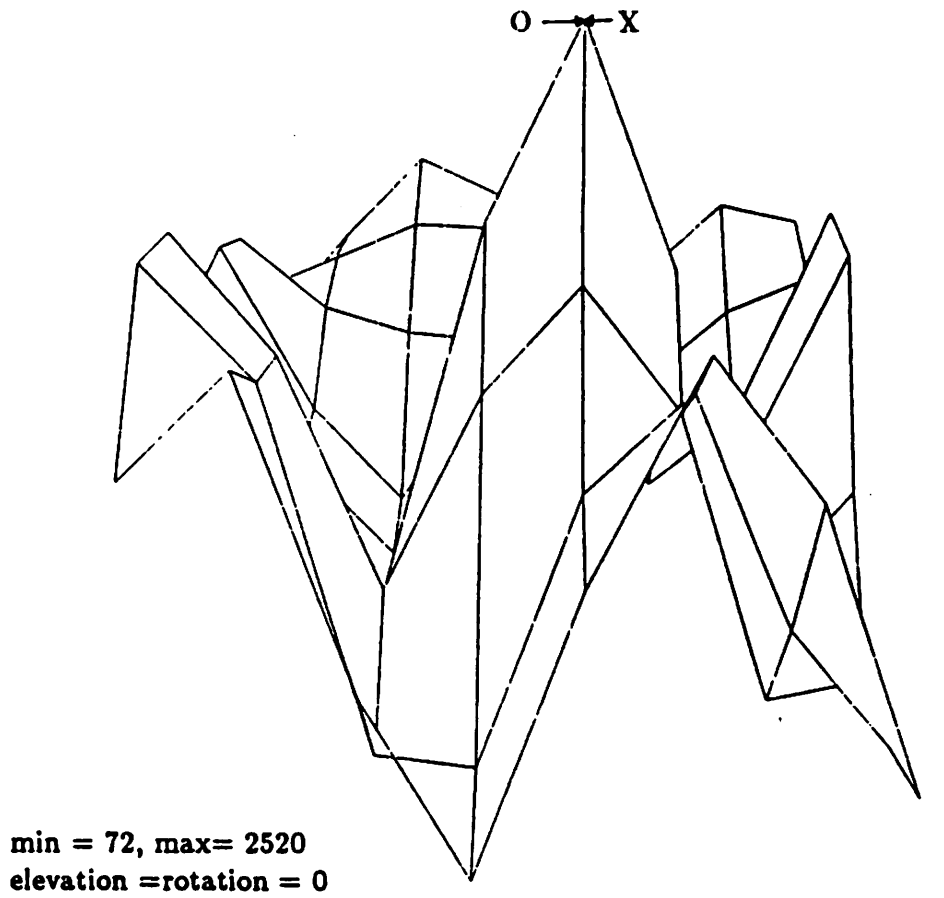

Figure 3: **The SSD surface at a corner point.**

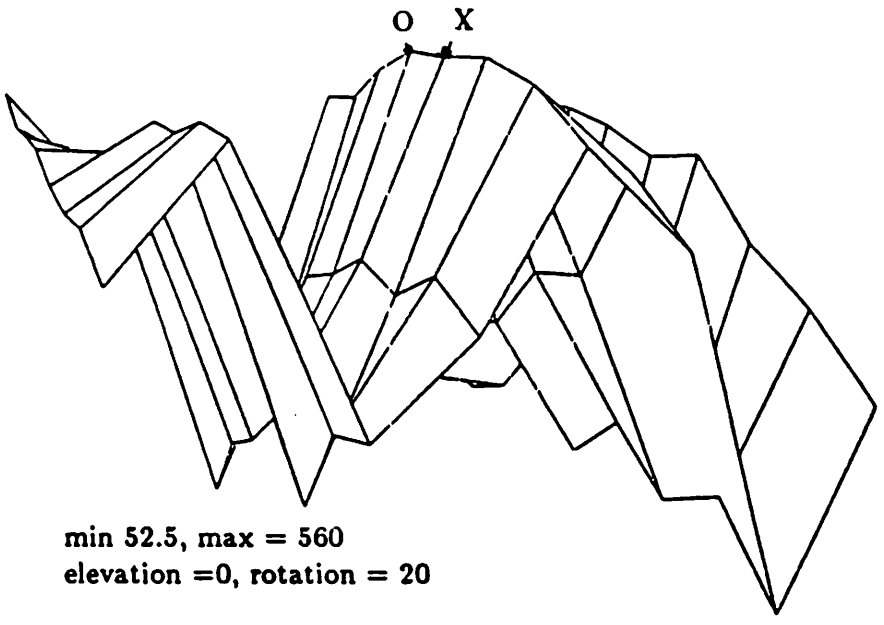

Figure 4: The SSD surface at an edge point.

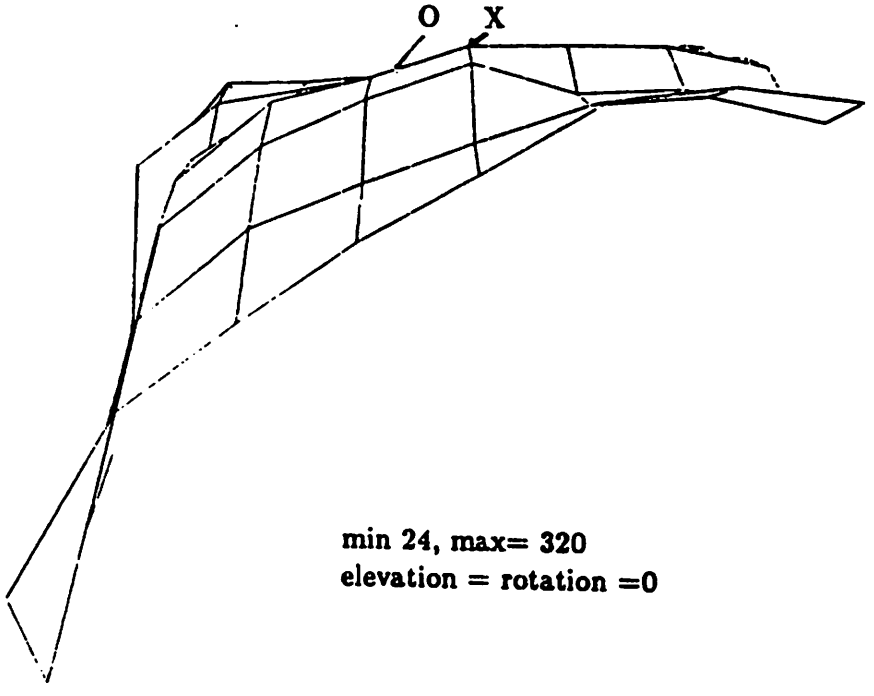

Figure 5: The SSD surface at a point in a homogeneous area.

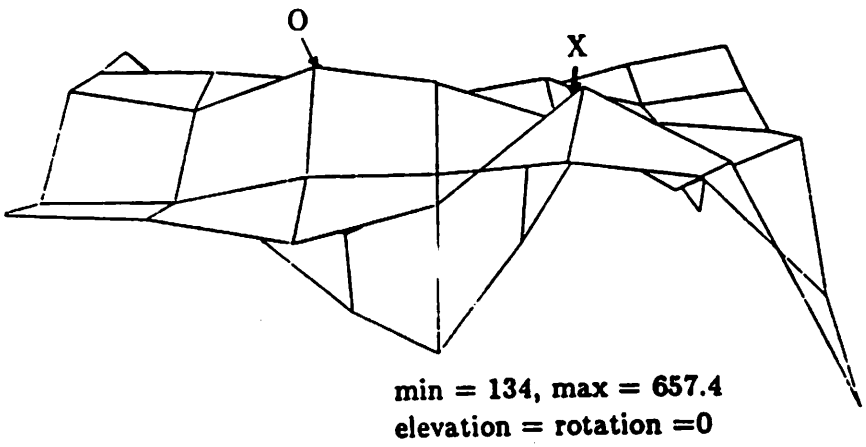

Figure 6: The SSD surface at a corner point which is occluded in the second frame.

The exact form of the normalization function was derived from the following considerations. The confidence measure is made proportional to the corresponding principal curvature according to the intuition that the reliability of the displacement should be proportional to the curvature of the SSD surface. Since a large value for S_{min} indicates an unreliable match due to occlusion, noise, or deformation of the image area, the confidence is inversely proportional to S_{min}. The presence of the term containing the curvature in the denominator is useful to maintain the confidence value in the range $(0, 1/k_3)$. If it is desired not to restrict the range of the confidence, k_3 can be set to zero. Finally, the constant term k_1 is used to maintain a finite value for the confidence measure when S_{min} tends to zero.

The computation of the principal curvatures involves knowing the second partial derivatives of the SSD surface along the coordinate directions. Given that we have a discrete set of SSD values, the following approach is used for this purpose. First, the 3×3 set of SSD values around the best match location are computed. Given this 3×3 set of SSD values, a quadratic surface fit can be extracted by using the best-least-square method of Beaudet [9]. This method involves computing weighted sums of the 3×3 values to obtain the various first- and second-order derivatives of a surface. The principal curvatures can be expressed as non-linear combinations of the second derivatives.

Discussion

Intuitively, the vectors $\hat{e}_{max}$ and $\hat{e}_{min}$ and the confidences c_{max} and c_{min} can be understood as follows: At a point along an edge in the image, the vector $\hat{e}_{max}$ will be approximately oriented in the direction normal to the edge, and $\hat{e}_{min}$ will be oriented parallel to the edge. At such a point, c_{max} will be large and c_{min} will be small. On the other hand, in an area of the image with small intensity variations, both the measures will be small, whereas at a point along a contour with high curvature, both will be high.

An important issue that was somewhat sidestepped in this section was the sensitivity to occlusion. Although we have so far assumed that a large value for S_{min} indicates a "false match" (i.e., a match does not *exist*), in general S_{min} can be large due to a variety of reasons. Some of these reasons are listed below:

1. The search area does not contain the true match, either because of an incorrect initial displacement or because of occlusion.

2. The template window contains a discontinuity in image flow. This arises at points near depth or motion discontinuities. In this case, since the window straddles the boundary, the intensity structure within the window varies between frames.

3. The magnitudes of rotation and/or the translation in depth are large or the image area undergoes non-rigid motion. In this case, the template window undergoes an area deformation, which violates the assumption of locally translational motion.

4. The SNR (signal-to-noise ratio) is low. In this case, the intensity values of corresponding areas in the two images differ due to the presence of noise.

In all the cases listed above, S_{min} will be large only if the local "spectral-energy" (i.e., the RMS value of the intensities in the template window) is large. A small value for the spectral energy suggests that the magnitudes of the intensity are small, i.e., the point is in a homogeneous intensity area; hence, all the values on the SSD surface will be uniformly small.

Finally, it should be noted that at a homogeneous occluded area, S_{min} will be small, even though the match estimate will usually be incorrect. As noted above, this arises due to the low spectral energy of the area. In these cases, the curvatures of the SSD surface may also be low, thereby making our confidence measures small. However, if the occluded homogeneous area straddles a textured area (where the spectral-energy is high), then the curvatures of the SSD surface may be large, because the displacements on one side of the

best match will have large SSD values, since they are a result of comparing a homogeneous area with a textured area. Therefore, although no match for the occluded point exists in the second image, the confidence measure may be large for some (incorrect) match *in the homogeneous area* of the second image. The estimated displacement is likely to equal the relative displacement between the textured area and the homogeneous area. Thus, there may be points in homogeneous areas that are occluded by textured areas, where our matching process may provide incorrect displacements with high values for the confidence measures. It should be noted that there are no solutions, or even approaches, to this type of problem in the literature on the measurement of motion.

During our empirical study of the SSD surfaces [7], we also observed that the shapes of the auto- and the cross-SSD surfaces are usually different for most occluded areas. Therefore, an additional clue for the presence of occlusion may be obtained by comparing the shape of these two surfaces. We expect to further develop these and other ideas for the detection of false matches during our own future research on the measurement of image motion.

3.5 Smoothness constraint

The problem of finding a smooth displacement field which approximates the displacement estimates computed at a discrete set of points by the local match process can been formulated as a minimization problem. That is, a vector field $\left\{\vec{U}\right\}$ is needed which minimizes a quadratic functional $E\left(\left\{\vec{U}\right\}\right) = E_{sm}\left(\left\{\vec{U}\right\}\right) + E_{ap}\left(\left\{\vec{U}\right\}\right)$ where E_{sm}, which is called the smoothness error, measures the spatial variation of $\left\{\vec{U}\right\}$ and E_{ap}, which is called the approximation error, measures how well $\left\{\vec{U}\right\}$ approximates the set of displacements provided by the matching process.

Intuitively, a displacement field can be considered "smooth" in an area of the image if its variation over the area is small. An example of a measure of the spatial variation of a displacement field is

$$\begin{aligned} E_{sm}\left(\left\{\vec{U}\right\}\right) &= \iint trace\left\{(\nabla U^T)(\nabla U^T)^T\right\}\, dx\, dy \\ &= \iint \left(u_x^2 + u_y^2 + v_x^2 + v_y^2\right)\, dx\, dy \end{aligned} \tag{1}$$

where $\{\vec{U}\}$ is the set of the displacement vectors $\vec{U}(x,y) = (u(x,y), v(x,y))$. The domain

of integration is usually the whole image. For notational convenience, we have used $\vec{U}$ to mean the vector $\vec{U}(x,y)$. The above formulation of a smoothness error is due to Horn and Schunck [26], who used this measure in a gradient-based approach for the computation of optical flow. Other examples of such measures will be discussed later in this section.

Let $\{\vec{D}\}$ be the set of estimates provided by the match process; these are represented in the local orthogonal basis $(\hat{e}_{max}, \hat{e}_{min})$, which denote the principal axes of the SSD surface. For a given displacement field $\left\{\vec{U}\right\}$, the approximation error is a weighted sum of the deviations of the components of the displacement vectors $\vec{U}(x,y)$ of the field along the basis directions from the corresponding components of the match estimates $\vec{D}(x,y)$. The weights are the confidences c_{max} and c_{min}. Mathematically,

$$\begin{aligned} E_{ap}\left(\left\{\vec{U}\right\}\right) &= \sum_{x,y}\Big[c_{max}(\vec{U}\cdot\hat{e}_{max} - \vec{D}\cdot\hat{e}_{max})^2 \\ &\quad + c_{min}(\vec{U}\cdot\hat{e}_{min} - \vec{D}\cdot\hat{e}_{min})^2\Big] \end{aligned} \tag{2}$$

Note that our formulation of E_{sm} implies that the space of admissible functions are a subclass of C_0 functions. The following alternate form which regards the admissible functions as a subclass of C_1 functions is also possible:

$$E_{sm:2,2} = \iint \left(u_{xx}^2 + 2u_{xy} + u_{yy}^2 + v_{xx}^2 + 2v_{xy}^2 + v_{yy}^2\right)\, dx\, dy \tag{3}$$

Although we have also implemented an algorithm based on this formulation (see [6]), most of the experiments have been based on the simpler formulation given above in equation 1. This is because results from early experiments with the two formulations did not show significant qualitative differences, while the solution to the second-order formulation given above requires more computational effort.

Solving the minimization problem

The two functionals E_{sm} and E_{ap} have been chosen in such a manner that under certain weak conditions there will always exist a unique solution for our minimization problem. In particular, it is easy to show that a unique minimum exists if there is at least one corner point in the image (both c_{max} and c_{min} are non-zero), and/or there are two points with different $\hat{e}_{max}$ vectors. A proof can be found in [7].

The most common approaches to solve the variational problem which has been formulated here are the finite difference method, a type of gradient-descent approach, and the finite-element method. For instance, Horn and Schunck [26], Glazer [21], and Nagel [34] all have used the finite-difference method. Hildreth [25] has used the conjugate-gradient approach, while Terzopoulos [38] has used the finite-element method for his surface reconstruction problem. We have also chosen the the finite-element method because it has a well-developed theory for the inclusion of known discontinuities in the field which is being determined.

The basic idea behind the finite-element method is the tessellation of the image plane using a set of elements with pre-defined shapes, and representing the field using piecewise polynomials defined over these elements. The order of the polynomials are determined according to the order of the derivatives involved in the error functional. A key requirement is that the discrete solution should converge to the true minimum as the element sizes tend to zero.

Terzopoulos has developed finite-element method algorithms for both the first- and the second-order smoothness constraints for the surface-interpolation problem. Since our variational problem can be regarded as a vector generalization of his scalar formulation, his solution methods can also be adapted. Since we have adapted Terzopoulos' approach for solving the variational problem, we do not discuss the mathematical details here. A clear description of such an analysis can be found in chapters 5 and 6 of [38]. Here, we simply describes the steps involved in our algorithmic implementation.

Computation of masks – In order to solve the discrete minimization problem, linear equations in the values at the nodes of a square-grid (which is in registration with the image-array) are derived. These equations are used to update the values $\vec{U} = (u, v)$ at a point in terms of its neighbors. In particular, solving the discrete problem can be shown to be the same as solving the following system of coupled equations:

$$\begin{aligned}(U - \overline{U}) \;&+\; c_{max}(U \cdot \hat{e}_{max} - D \cdot \hat{e}_{max})\hat{e}_{max} \\ &+\; c_{min}(U \cdot \hat{e}_{min} - D \cdot \hat{e}_{min})\hat{e}_{min} = \vec{0}\end{aligned} \tag{4}$$

where for each point on the grid, $\overline{U}$ is a weighted average of the displacements of its neighbors. For the first-order smoothness constraint the weights are distributed as follows:

$$\frac{1}{4} \times \begin{array}{ccc} & 1 & \\ 1 & 0 & 1 \\ & 1 & \end{array}$$

Relaxation algorithm – There are a number of numerical methods for solving the system of coupled linear equations described above. One of the simplest methods is the Gauss-Seidel relaxation algorithm. This is an iterative process, where during each iteration the value of U at each point in the image is solved in terms of the values of its neighbors.

The iterative update equation for the displacement field smoothing problem is,

$$\begin{aligned} U^{n+1} = \overline{U}^n \; &+ \; \frac{c_{max}}{c_{max}+1}((D-\overline{U}^n)\cdot\hat{e}_{max})\hat{e}_{max} \\ &+ \; \frac{c_{min}}{c_{min}+1}((D-\overline{U}^n)\cdot\hat{e}_{min})\hat{e}_{min} \end{aligned} \tag{5}$$

where the superscripts denote the number of the iteration.

The updating scheme described above has the following geometric interpretation: U^{n+1} is a point in the (u,v) space which is a combination of $\overline{U}^n$ and D. The manner in which this combination takes place is illustrated in Figure 7, where the displacements have been represented in a cartesian coordinate system with its axes parallel to $(\hat{e}_{max}, \hat{e}_{min})$. Since $c_{max} \geq c_{min}$, the location of U^{n+1} will always be on or above the line joining D and $\overline{U}^n$. In particular, it can be seen that U^{n+1} always will be within the triangle shown in the figure, moving towards the line joining D and $\overline{U}^n$ as c_{min} gets closer to c_{max}.

The two key parameters are $c_{max}/(1+c_{max})$ and $c_{min}/(1+c_{min})$, which vary between 0 and 1, as c_{max} and c_{min} vary between 0 and ∞. When $c/1+c$ (where $c = c_{max}$ or c_{min} appropriately), is close to zero, the updated value is close to the average of the neighbors, whereas when it is close to 1, the updated value is close to the initial local displacement estimate. The function $c/1+c$ rises rapidly and approaches its maximum value 1, so that even a small value of c (e.g., 10) orients the updated value strongly $\left(\frac{c}{1+c} = 0.91\right)$ towards the initial local estimate. For our experiment, the choice of the normalization parameters for the confidence measures was based on this observation concerning the behavior of the updating algorithm.

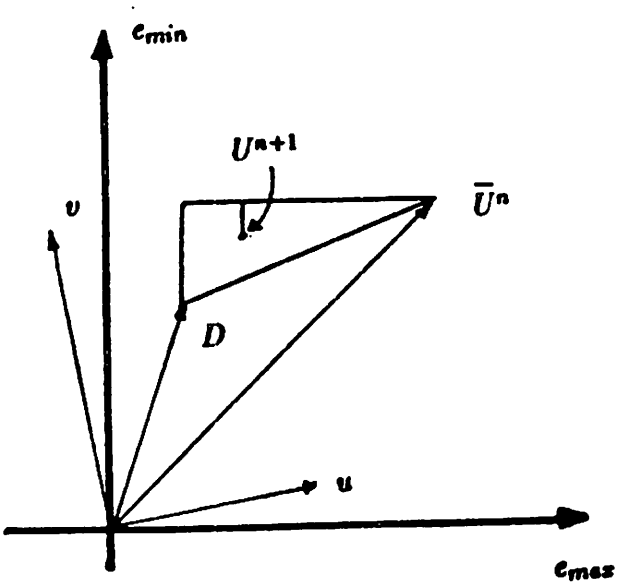

Figure 7: A geometric interpretation of the relaxation process

Finally, note that during the projection of the displacements to the next finer-level, the values will be rounded to the nearest integer value; hence, it is not necessary to wait until the complete convergence of the smoothing algorithm. The relaxation process can be stopped when the rounded-off values of the displacements do not change during an iteration. In practice, we found that usually 10 iterations were sufficient to achieve this condition.

4 Experimental Results

This section describes the results applying the ideas described in this paper to two pairs of real images. The two experiments are called the "dinosaur image experiment" and the "hallway scene experiment".

The key parameters involved in the computation of our confidence meaures are the normalization coefficients k_1, k_2, and k_3. For our experiment, we chose $k_1 = 150, k_2 = 1$ and $k_3 = 0$. As noted in section 3.4, k_1 is an overall scaling factor, k_2 controls the influence of S_{min}, and k_3 is useful to restrict the range of the confidence values. The choice of $k_3 = 0$ simply removes any restrictions on the range of the confidences. The choice of $k_2 = 1$ means that the influence of S_{min} is of the same order as that of the curvatures. Our choice of $k_1 = 150$ was based on the empirical observation that the mean values of C_{max} was between 100 and 200. Therefore, barring the effects of S_{min}, on the average c_{max} will be approximately 1, and the factor $c_{max}/1 + c_{max}$ will be approximately $1/2$. This means that on the average, the local displacement estimate and the weighted average of the neighbors' estimates have equal effect during each iteration of the relaxation process.

4.1 The dinosaur image experiment

The input images used in this experiment are the two 128×128 resolution images shown in Figure 8. The scene consists of a toy-dinosaur, a toy-chicken in the background, and a tea-box in the foreground, all of which rest on a table-top which has a grid-pattern on it. The toy-chicken, which is somewhat hard to see in the images shown here, is behind the neck area of the toy-dinosaur. The 3-D motion between the two frames consists of a translation of the camera to the right along with a leftward rotation of 1.5° about the vertical axis (in order to bring the scene back into view), as well as an independent 4.2° anti-clockwise rotation of the dinosaur about the optical axis. The magnitudes of the image displacements induced by these 3-D motions are as follows: the toy-chicken appears to have

moved rightwards by about 3 to 5 pixels; the tea-can appears to have moved leftwards by about 4 to 6 pixels; the grid pattern on the floor is almost stationary; and the points on the surface of the dinoasur appear to have moved by about 7 to 10 pixels. This scene is of interest for the obvious reason that it contains a distinct and prominent independently moving object besides a complex camera motion.

Figure 9 shows the displacement fields at the four levels of the pyramid computed by the hierarchical matching process *without* smoothing, while Figure 10 displays the displacement field at the same four levels produced by the hierarchical algorithm *with* smoothing. Figure 11 displays the displacement field at the finest resolution superimposed on the first frame. Figure 12 displays the confidence measure c_{max} at the four levels with the direction vectors $\hat{e}_{max}$ superimposed, as well as the confidence measure c_{min}.

Qualitatively, Figure 11 shows that the algorithm performs remarkably well in this real image containing complex motion. Note that while the toy-chicken and the tea-box undergo the same 3-D relative motion with respect to the camera, (i.e., a translation parallel to the image plane combined with a rotation about the vertical axis), the movement of their images appear to be in opposite directions. This is because, while the leftward rotation of the camera induces a rightward image-flow in both the regions, the effect of the compensatory rightward translation is greater on the image of the tea-box, since it is closer to the camera. Figure 11 shows that our algorithm has correctly determined the image-displacements of these two objects. It is also clear that the independent movement of the dinosaur has been successfully computed.

The expected behaviors of the confidence measures at corners, edges, and homogeneous areas are confirmed by the displays in Figure 12. In order to make these behaviors more explicit, we attempted to classify the image pixels as corners, edges, or homogeneous areas according to the following criteria:

- if $c_{min} > 0.5$ the pixel is classified as a corner,
- if $\frac{c_{max}}{c_{min}} > 100$ and $c_{max} > 1$, the pixel is classified as an edge, and
- if $c_{max} < 0.5$ and $\frac{c_{max}}{c_{min}} < 5$ the pixel is classified as a point in a homogeneous area.

This classification is shown in Figure 13.

The improvement obtained by the smoothing process is easily seen by comparing the results shown in Figures 9 and 10. Note that at the two coarsest resolutions, the dis-

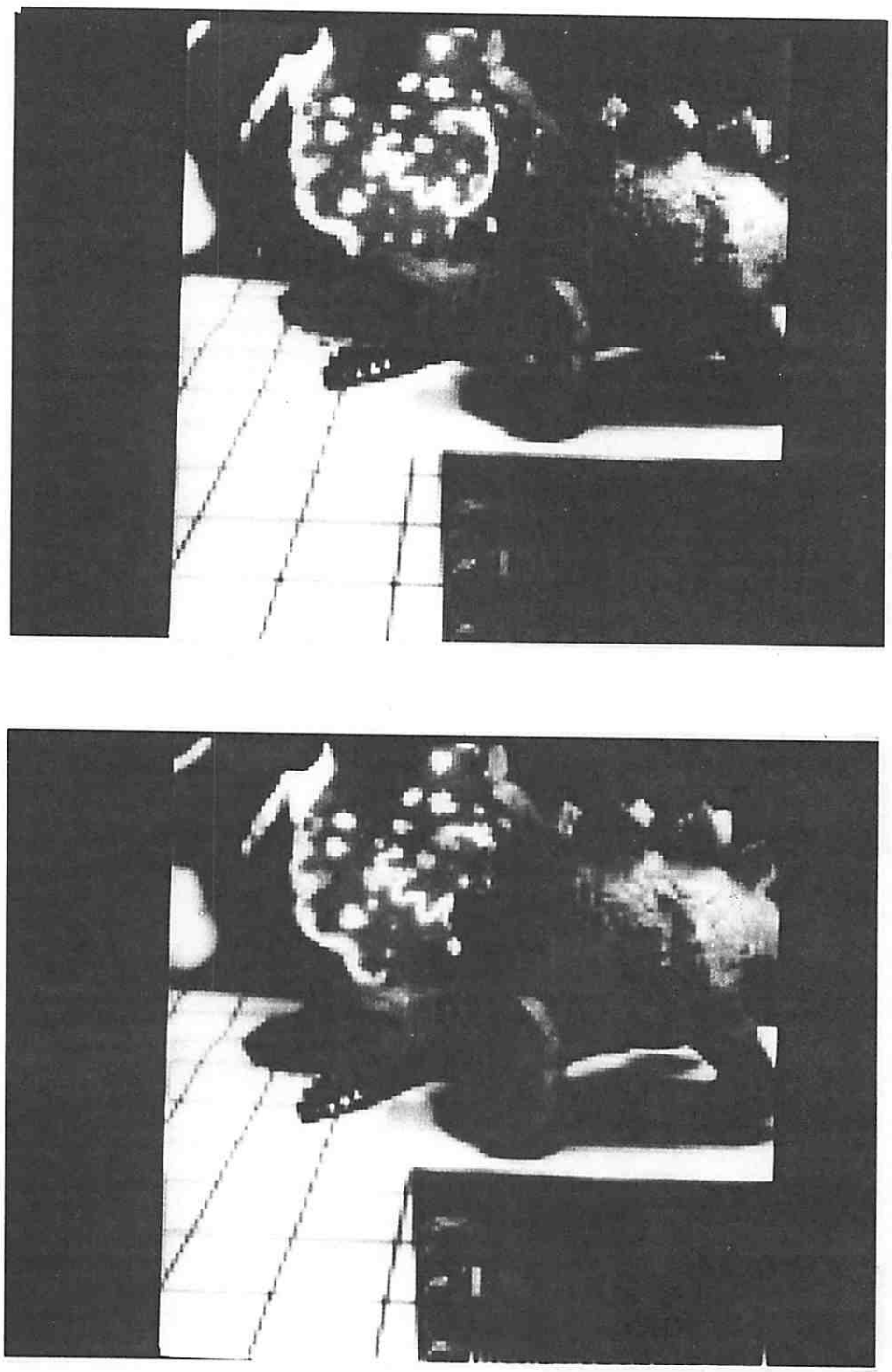

Figure 8: **The dinosaur-image experiment – input images**

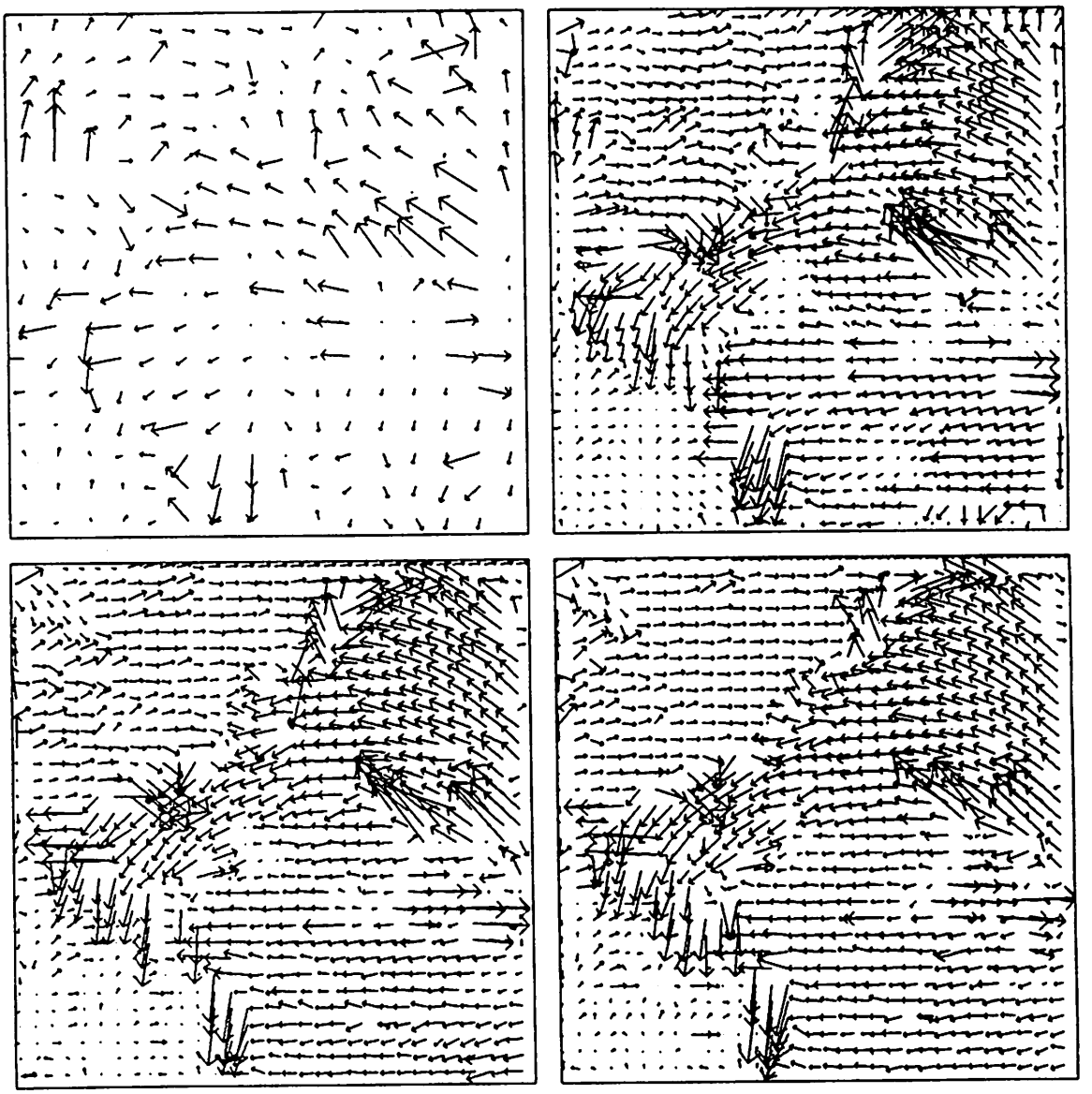

Figure 9: The dinosaur-image experiment – unsmoothed results
The results of the dinosaur-image experiment without smoothing and with 5×5 Gaussian template windows. The top left and right quadrants contain results from levels 4 and 5 respectively, while the bottom left and right quadrants contain the results from levels 6 and 7. In order to enhance visibility only a 32×32 sample of the displacements have been shown at levels 6 and 7.

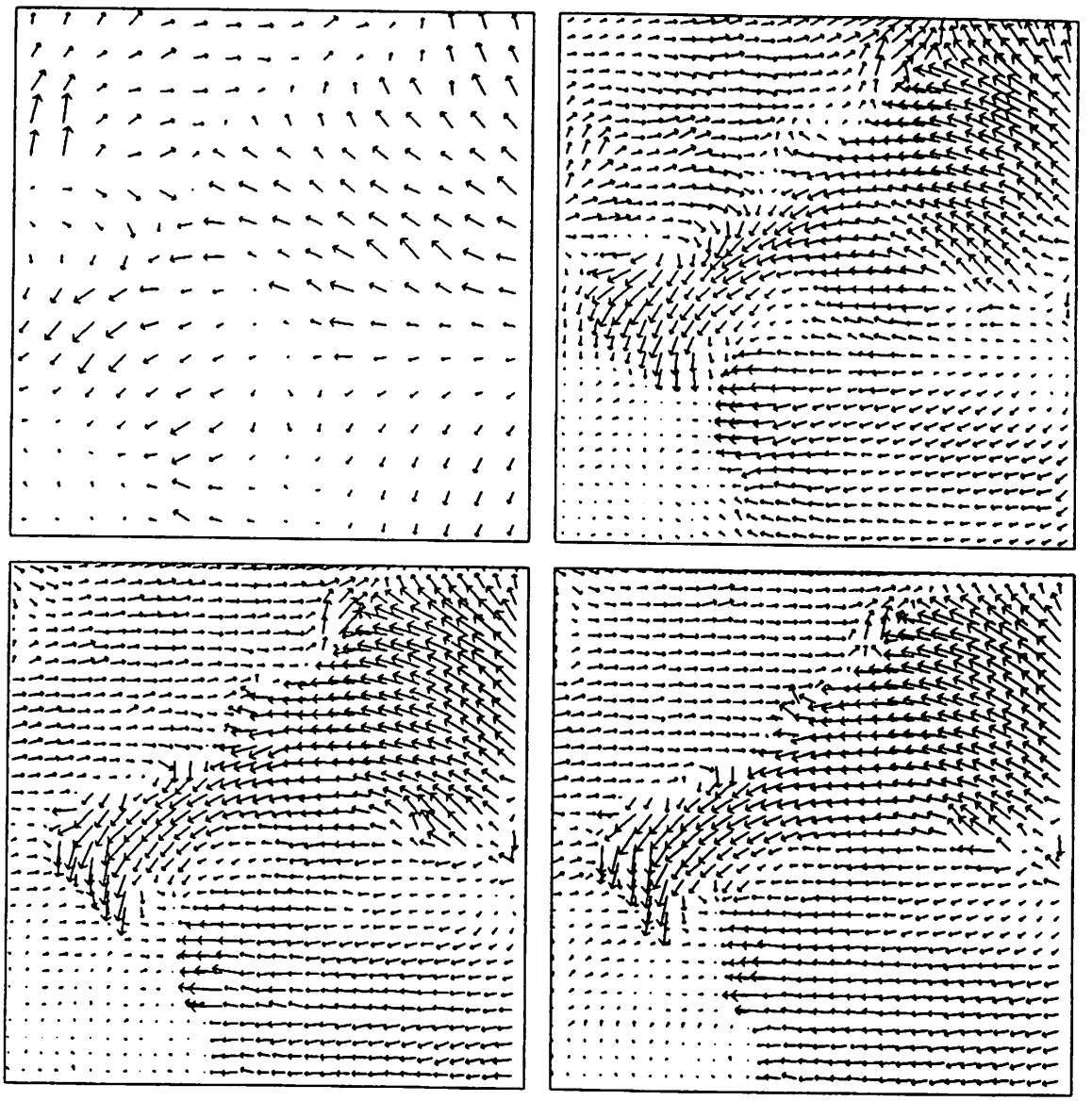

Figure 10: The dinosaur-image experiment – smoothed results
The results of the dinosaur-image experiment with smoothing and with 5×5 Gaussian template windows. Once again, the top left and right quadrants contain results from levels 4 and 5 respectively, while the bottom left and right quadrants contain the results from levels 6 and 7. Also, only a 32×32 sample of the displacements have been shown at levels 6 and 7.

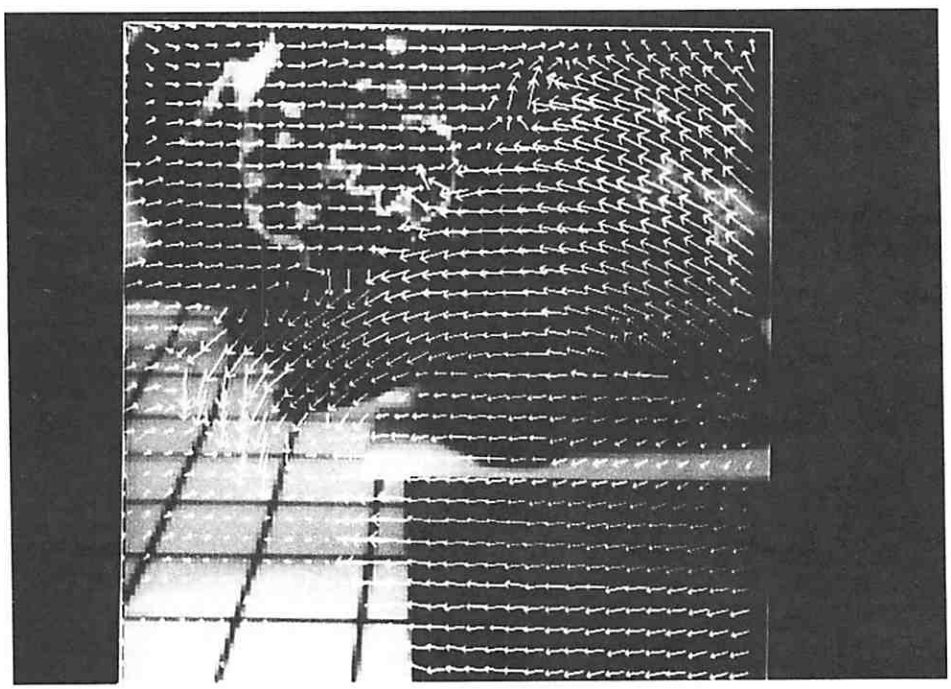

Figure 11: **The dinosaur-image experiment – finest level results**
The smoothed displacement vector field at the finest level for the dinosaur-image experiment superimposed on the first input frame. In order to enhance visibility only a 32×32 sample of the displacements have been shown.

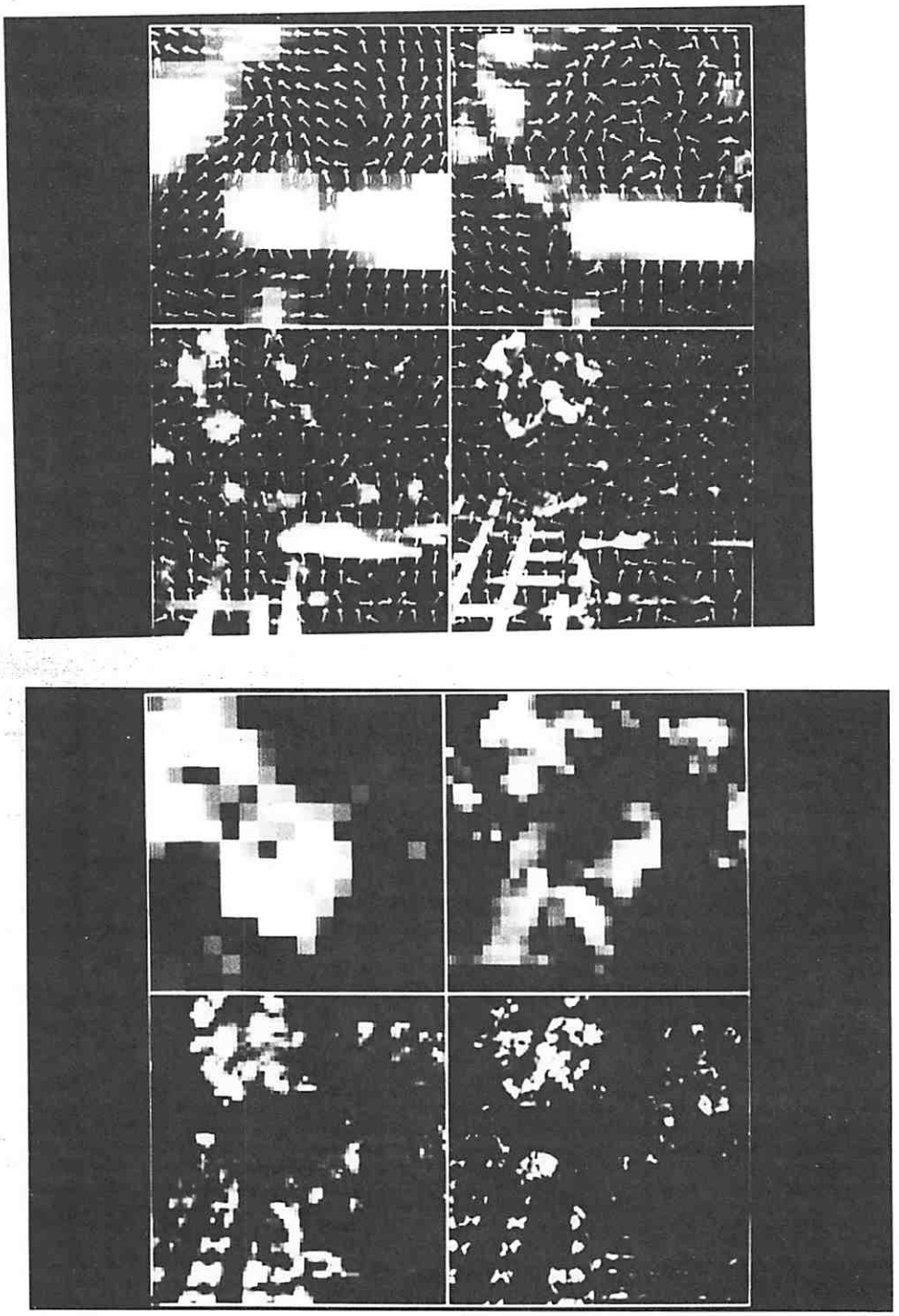

Figure 12: **The dinosaur-image experiment – confidence measures**
The confidence-measures computed in the dinosaur-image experiment with smoothing. The confidence measure are shown at the four finest levels. The top figure shows c_{max} and samples of $\hat{e}_{max}$, while the bottom figure displays c_{min}.

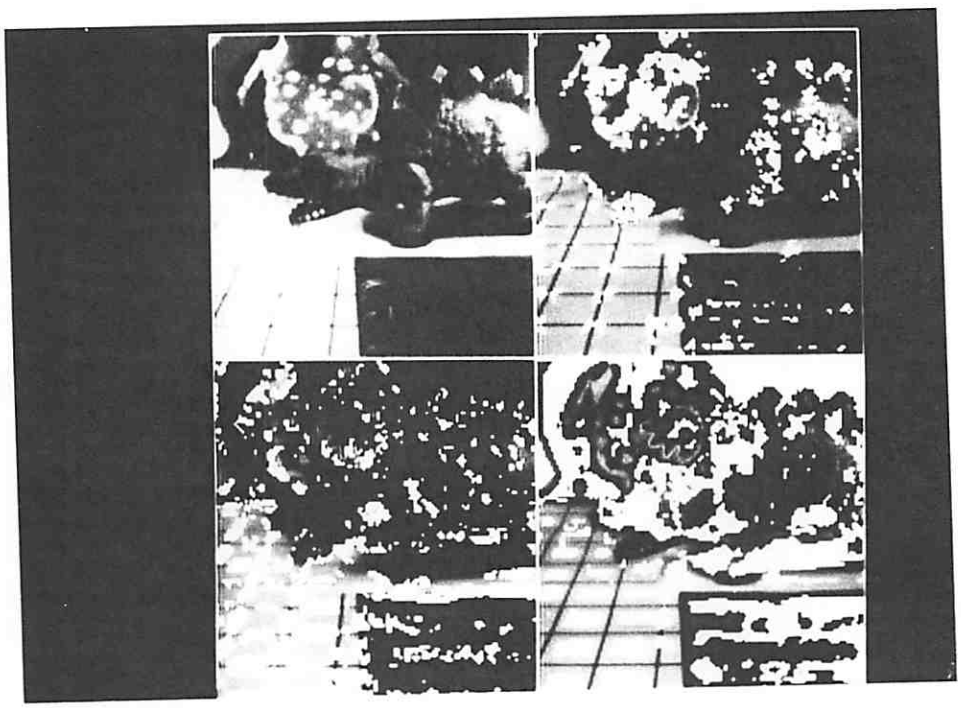

Figure 13: **The dinosaur-image experiment – pixel classification**
The classification of pixels as corners, edges, or homogeneous areas. The top left quadrant contains the input image. In the images shown in the top-right, bottom-left, and bottom-right quadrants, the corners, edges, and the homogeneous areas have respectively been highlighted.

placement field has been smoothed across surface and object boundaries, whereas at the finer-levels there are sudden changes near such boundaries. This is due to the use of overlapped pyramid projection strategy, as well as the fact that the input images contain significant contrast at high-frequencies. Hence, the finer-level matching processes were able to correct some of the errors made at the coarser-levels. In particular, note that the boundary between the chicken and the dinosaur has been maintained during the finer-level smoothing processes, due primarily to the proper decoupling via the confidence measures.

Although the area on top of the dinosaur is part of the background, the vectors in that area seem to be influenced by the motion of the dinosaur. Similarly, the area of the floor just left of the tea-box has displacements that are obviously incorrect. This is because both these areas are somewhat homogeneous, and are adjacent to areas containing high-contrast information. In addition, parts of the floor have been occluded, or are near the occlusion boundary, and therefore do not have reliable local estimates. Hence, the more reliable neighboring vectors have been propagated by the smoothing process to these areas with unreliable local estimates. Such problems due to occlusion are pervasive in the field of motion analysis.

Note that near the top of the tea-box, one of the lines from the grid-pattern is visible in both the frames, although its displacement is incorrect. This error occurs because the vertical line on the floor is adjacent to the occlusion boundary, and the intensity structure of its neighborhood undergoes significant changes between the two frames. The problems here have been made more severe by the use of a coarse-to-fine control strategy, since at low-frequencies the area affected by the occlusion increases; it appears that even the use of the overlapped pyramid projection strategy has not corrected these errors. However, as illustrated in Figures 12 and 13 the confidences associated with the incorrect displacements on the line are small. In particular, note that in Figure 13, this line has been classified as a homogeneous area. This is because our simple classification scheme does not discriminate between low-confidences due to homogeneous intensity structures and due to occlusion. As suggested in section 3.4, a comparison of the auto- and the cross- SSD surfaces would reveal that this area contains a linear structure, and is either occluded or is adjacent to an occlusion boundary.

The area of the floor just below the nose of the dinosaur also has incorrect displacements. In this case, however, the problem is not due to the smoothness constraint. Instead, it arises because the grid-pattern on the floor is periodic, and the difference in the displacement of the nose of the dinosaur and the floor is approximately equal to the period of the

grid-pattern. Hence, at the coarse-levels of processing, the grid-pattern near the nose of the dinosaur appears to have moved one-period, whereas its actual image motion is much smaller (almost zero). This may be a harder problem to solve by a low-level two-frame matching technique, because all displacements which are multiples of the period of the grid-pattern are equally valid. In order to resolve between them, either higher-level texture-based grouping processes, or constraints involving the temporal coherence of the movement may be necessary. For a discussion of the mechanisms for incorporating the temporal coherence assumption, refer to [7].

4.2 The hallway scene experiment

The input images for this experiment are shown in Figure 14. These 256 $\times$ 256 pixel resolution images were produced at the UMass Computer Vision Laboratory. Once again all image motion is due to a camera undergoing translational motion. The reason for choosing this image-pair is the presence of the many long linear-structures in the image. Since the confidence measure separates such areas from corners and homogeneous areas, it is interesting to study its behavior in this experiment.

Due to length considerations of this paper, we only show the results of the algorithm with smoothing at the finest level. Figure 15 shows these displacements superimposed on the first image frame. Just as in the case of the dinosaur image experiment, we attempted to classify the pixels of the image as corners, edges, and homogeneous areas. The criteria used are the same as those described in the dinosaur image experiment. Figure 16 displays the results of this classification.

In order to further illustrate the accuracy of our computations, and to confirm our statement regarding the normal and the tangential components, we superimposed our displacement vectors on the set of lines obtained from the two input images by a line-extraction and grouping algorithm developed by Boldt [10,44]. Figure 17 displays the lines extracted from the two images; the lines extracted from the first frame are shown as thick dark lines, while the thinner lines are those obtained from the second frame. Only lines whose associated contrast is greater than 15 grey-levels and which are longer than 7 pixels have been shown. In Figures 18 and 19, we have superimposed our displacement vectors on the sets of lines obtained from the two-frames in a 90×90 pixel area, which is marked in Figure 17. Figure 18 displays the *unsmoothed* displacement vectors for a sample of pixels which lie on the lines belonging to the first frame, while Figure 19 shows the *smoothed* displacement vectors for the same sample of pixels.

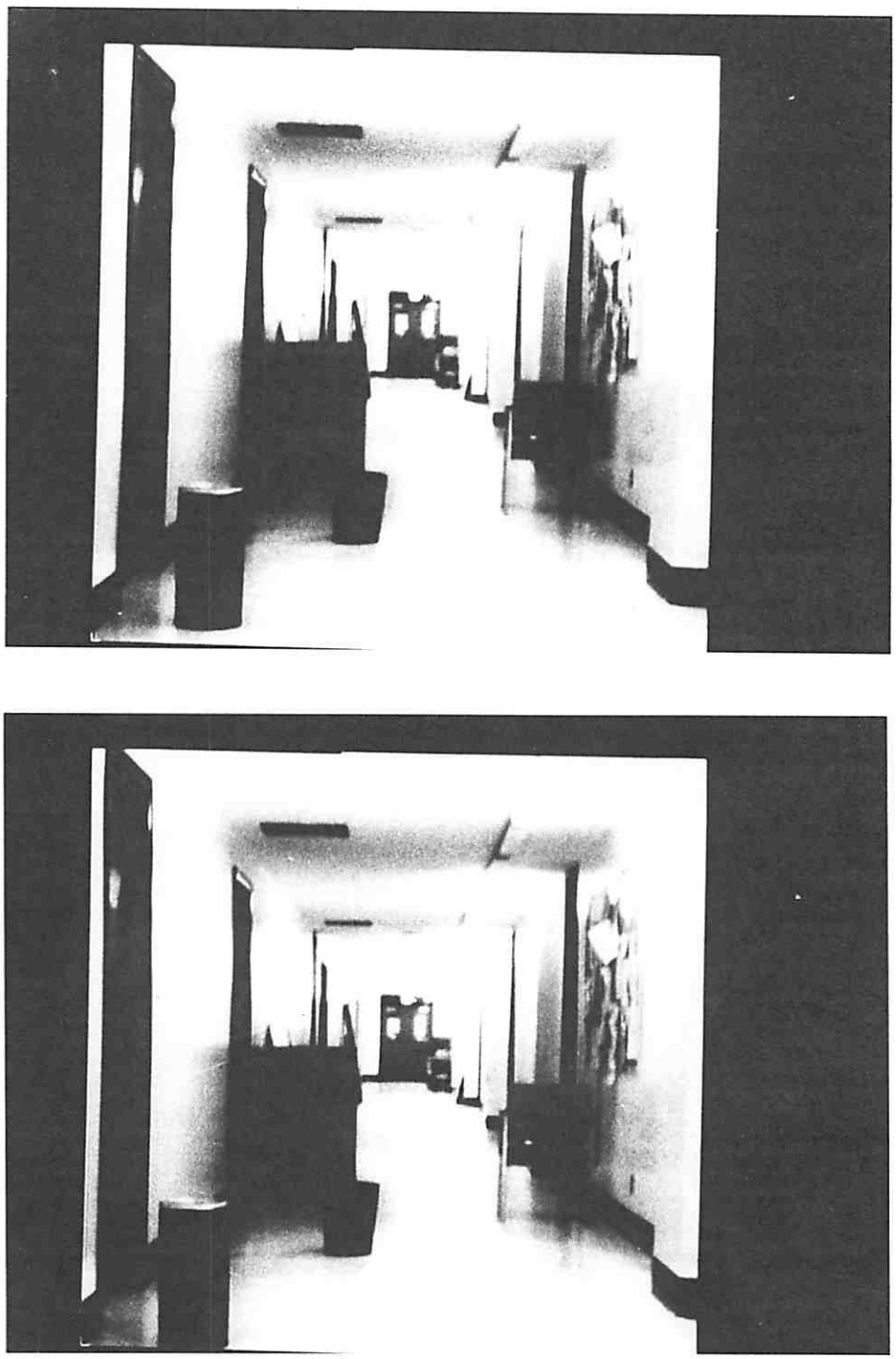

Figure 14: **The hallway-scene experiment – input images.**

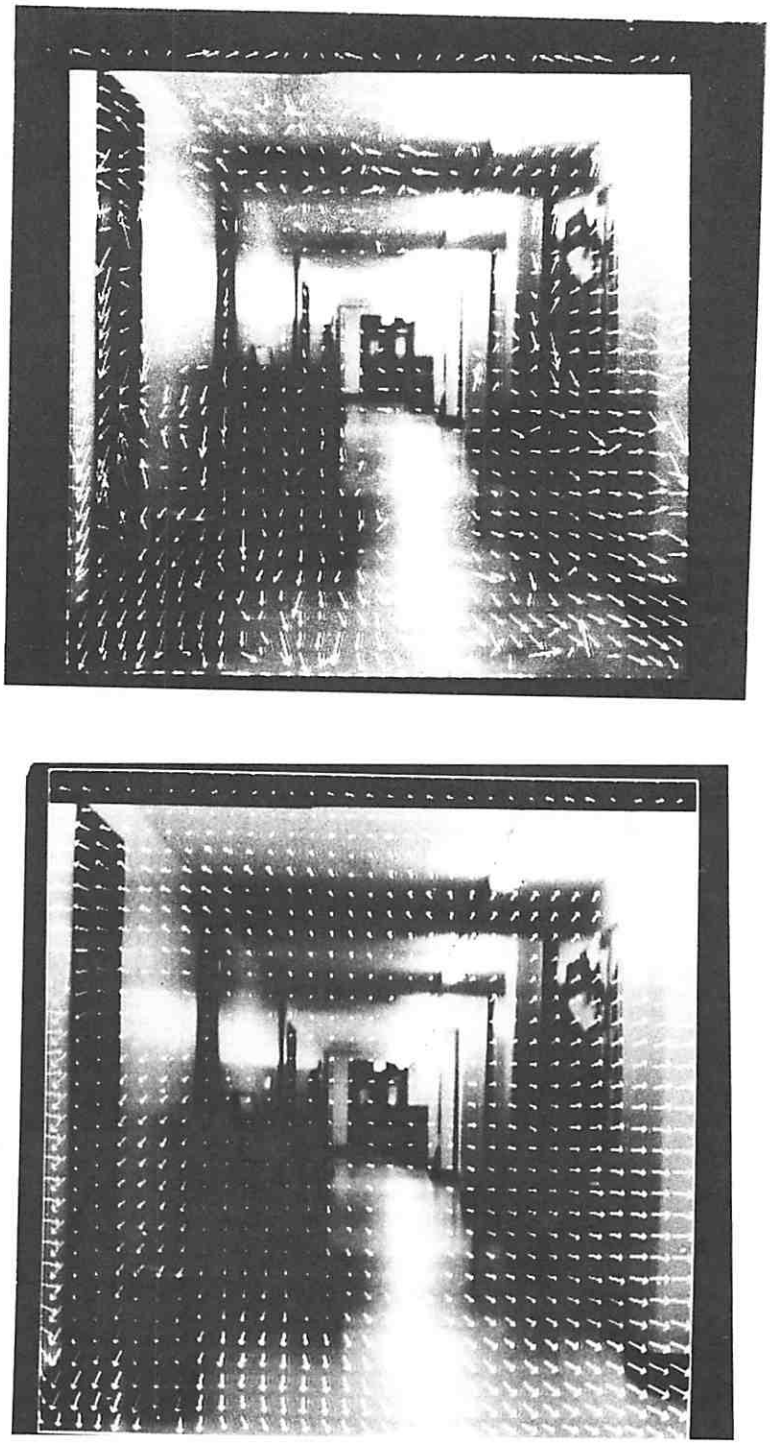

Figure 15: **The hallway-scene experiments – finest level results**
The finest-level displacements for the hallway-scene experiment superimposed on the first input frame. (a) without smoothing, and (b) with smoothing. In order to enhance visibility only a 32×32 sample of the displacements have been shown.

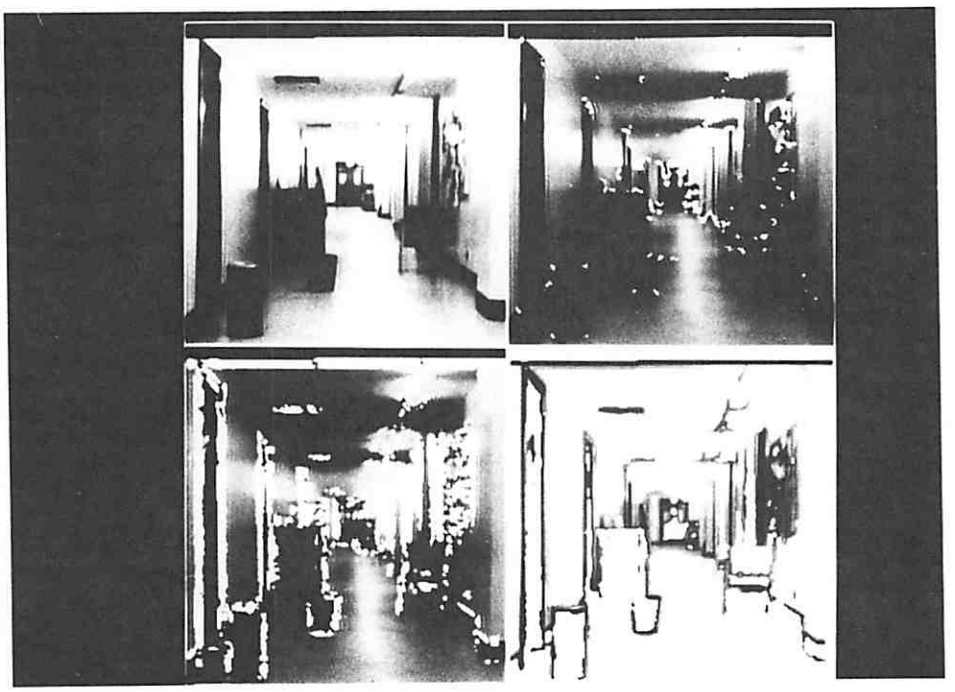

Figure 16: **The hallway-scene experiment – pixel classification**
The hallway-scene experiment: The classification of pixels as corners, edges, or homogeneous areas. The top left quadrant contains the input image. In the images shown in the top-right, bottom-left, and bottom-right quadrants, the corners, edges, and the homogeneous areas have respectively been highlighted.

Figure 17: **The hallway-scene experiment – lines**
The hallway-scene experiment: The lines extracted by Boldt's algorithm from the two input frames. The thick lines are those from the first frame, while the thin lines are from the second input frame. The area within the rectangular box in the upper left corner of the image will be more closely examined in the next two figures.

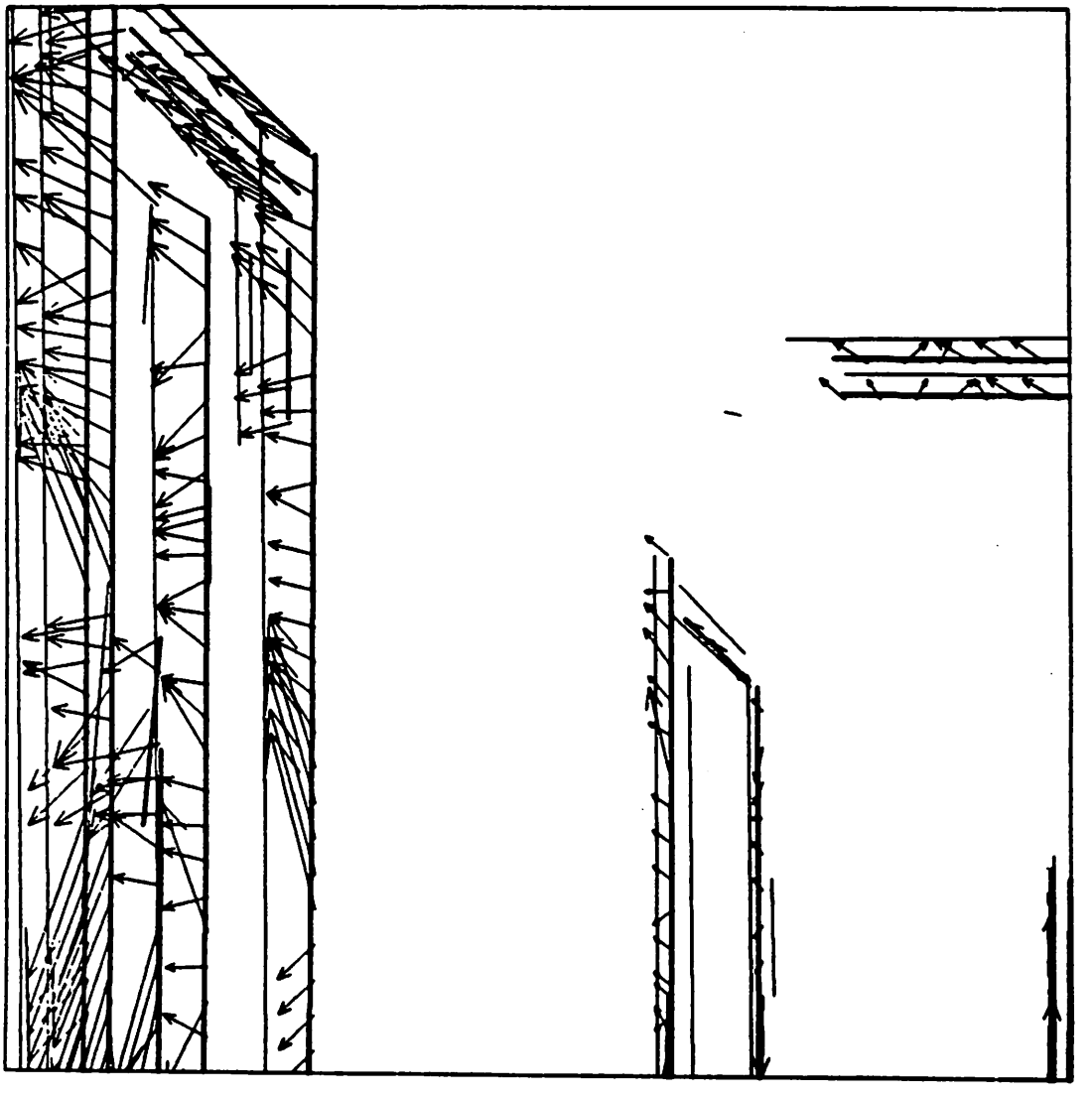

Figure 18: **The hallway-scene experiment – lines with unsmoothed displacements**
The hallway-scene experiment: The unsmoothed displacements superimposed on the lines extracted by Boldt's algorithm.

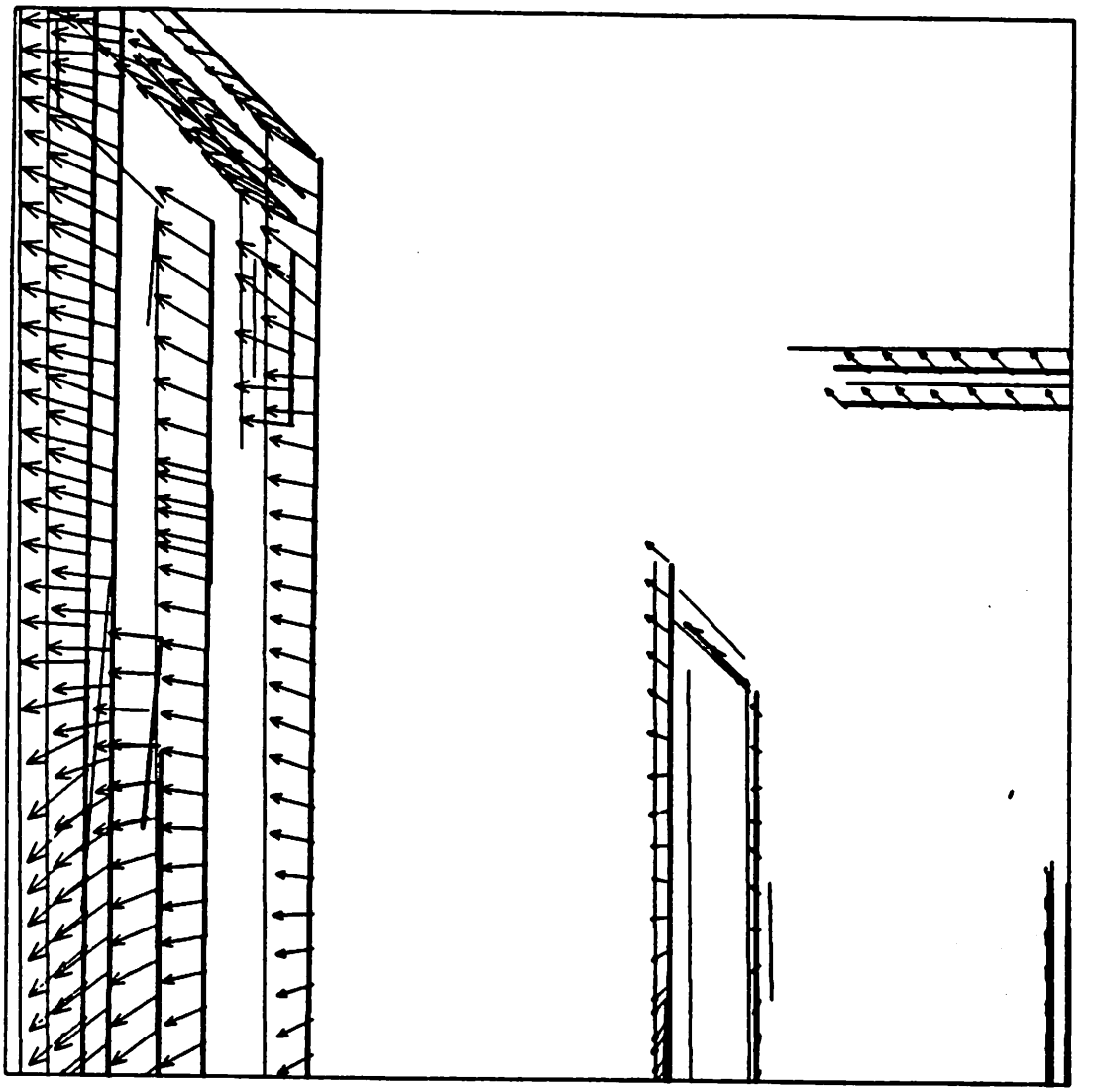

Figure 19: The hallway-scene experiment – lines with smoothed displacements
The hallway-scene experiment: The smoothed displacements superimposed on the lines extracted by Boldt's algorithm.

It is obvious from Figure 19, that the lines are correctly matched by our displacement vectors. From Figure 18, it is clear that the normal-components of the unsmoothed vectors are also correct, whereas their tangential components are often incorrect. Finally, the remarkable consistency of the results obtained by Boldt's line-extraction algorithm and our matching algorithm suggests that it may be possible to combine them to extract stable line tokens from image sequences. This idea is currently being pursued at the UMass Computer Vision Laboratory.

5 Relationship to the Gradient-Based Techniques

In this section, we describe the general principles underlying the gradient-baed techniques for computing image velocity fields, describe two specific gradient-based algorithms and show that both these techniques are consistent with our framework. Further, we will also show that there is a clear mathematical relationship between the two gradient-based techniques and our matching algorithm.

5.1 The gradient-based techniques – an overview

Almost all the techniques for measuring the instantaneous image-velocities use the gradient-based approach. This approach is based on the assumption that the intensity of light reflected by a point on an environmental surface and recorded in the image remains constant during a short time interval, although the location of the image of that point may change due to motion. This assumption leads to the following equation, which is called the *intensity constraint*:

$$|\nabla I| U^{\perp} = -I_t, \tag{6}$$

where $|\nabla I|$ is the magnitude of the intensity gradient-vector $\vec{\nabla} I = (I_x, I_y)$, I_t is the temporal derivative of the intensity function, and $U^{\perp}$ is the component of the image velocity $\vec{U}$ parallel to $\vec{\nabla} I$. The other component U^T, along the direction perpendicular to $\vec{\nabla} I$, is unspecified by this constraint. Since the orientation of the intensity gradient vector is normal to the direction of the "edge" at a point, $U^{\perp}$ and U^T are respectively called the "normal-flow" and the "tangential-flow" components of the edge. The lack of information regarding the tangential-flow component is known as the "aperture problem" [39].

Since the intensity constraint specifies only one component of the image-velocity at a point, an additional constraint is necessary to completely determine the velocity vector;

this is usually given in the form of a smoothness constraint on the velocity field [25,26,34]. For this paper, Horn and Schunk's formulation [26] and Nagel's formulation [34] are of particular interest, because they are respectively used in the multi-resolution gradient-based algorithms of Glazer [21] and Enkelmann [17]. These two formulations of the smoothness constraints are discussed in detail within the descriptions of the hierarchical techniques given below. It should be noted that both techniques apply the intensity constraint for the computation of displacement. Such an approximation of velocities by displacements is reasonable because of the hierarchical processing schemes used, in which all displacement measurements are small compared to the scale of image-intensity variations.

Enkelmann's approach

Enkelmann [17] uses the *low-pass pyramid* transform described by Crowley and Stern [16] to create a set of Gaussian low-pass filters. After the construction of a low-pass pyramid from each image, the processing begins at a particular coarse level; the description of his technique does not specify how this level is chosen. At the coarsest level, the initial displacement field consists of vectors of zero length. At all other levels the initial displacement field is determined by projecting the field computed at the adjacent coarser level. The projection process involves a bi-linear interpolation of the displacement vectors in a small neighborhood of the field at the coarse level.

Within each level, the process of refining the initial displacements is based on Nagel's gradient-based approach [34]. In this approach, an area around each pixel in the image is shifted according to the initial displacement vector at that pixel. The refinement to the initial displacement field is computed by minimizing the functional $E = E_{int} + \alpha^2 E_{sm}$ where E_{int} is a formulation of the intensity constraint mentioned above, and E_{sm} represents the smoothness assumption, and measures the spatial variation of the displacement field. Mathematically,

$$E_{int} = \iint dx\,dy\ (I(x,y) - J(x+u, y+v))^2 \tag{7}$$

and

$$E_{sm} = \iint dx\,dy\,trace\left[(\nabla U^T)^T W (\nabla U^T)\right], \tag{8}$$

where I is the intensity function of the first image, J is the intensity function of the second image, and W is a weight matrix which depends on the spatial derivatives of the image intensity function I. The inclusion of W has the effect of allowing the free propagation of

the smoothness assumption along image-contours, while restricting its propagation across the contours.

By using the Euler-Lagrange equations, and ignoring the second-order terms of (u, v), as well as the third- and higher-order spatial derivatives of the intensity function, the functional minimization problem is transformed to that of solving the following differential equations:

$$AU + b - \alpha^2 \begin{bmatrix} trace\{W\nabla\nabla u\} \\ trace\{W\nabla\nabla v\} \end{bmatrix} = 0, \tag{9}$$

where

$$A = (\nabla I)(\nabla I)^T + \bar{x^2}(\nabla\nabla I)(\nabla\nabla I)^T, \tag{10}$$

and

$$b = \Delta I(\nabla I) + \bar{x^2}(\nabla\nabla I)(\nabla\Delta I). \tag{11}$$

The $\nabla\nabla$ operator represents the matrix of second derivatives, e.g.,

$$\nabla\nabla I = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix},$$

and $\bar{x^2}$ denotes the size of a small image-window around a point (x, y) which represents that point. Note that in the limit, when the time interval between the two frames tends to zero, the displacements in the above differential equation can be replaced by the corresponding image-velocities, provided ΔI is replaced by I_t, the temporal intensity derivative.

By using the finite-difference approach, the differential equations are further transformed into a large sparse system of linear equations. These linear equations are then solved using a multi-resolution relaxation approach. The details of the relaxation process are not relevant for the purposes of this paper, although they may be important for an efficient implementation.

Glazer's approach

Glazer also uses a Gaussian low-pass pyramid representation of the input images and employs a hierarchical version of the Horn and Schunck approach. However, the exact algorithm for the construction of the pyramid is different; Glazer uses Burt's Gaussian-pyramid transformation described in [13]. After the construction of the low-pass pyramids, the processing begins at a coarse level at which the magnitudes of the displacements are expected to be less than a pixel. A coarse-to-fine control-strategy is used; the projection

of the displacements between adjacent levels is structured via the quad-tree connectivity, wherein each pixel at a coarse level is regarded as the "parent" of four pixels at the adjacent finer level. Each "child" uses the displacement of the parent pixel as its initial displacement.

As in Enkelmann's approach, a window around each pixel in the first image is shifted according to the initial displacement at that pixel. The refinement process also consists of minimizing the sum of two functionals E_{int} and E_{sm}, which represent the intensity constraint and the smoothness assumption. Glazer defines the two errors as

$$E_{int} = \iint dx\, dy\, |\nabla I|^2 (U^{\perp} - V^{\perp})^2, \tag{12}$$

and

$$E_{sm} = \iint dx\, dy\, (\nabla U^T)^T (\nabla U^T). \tag{13}$$

As before, $U^{\perp}$ is the component of the displacement vector $\vec{U}$ parallel to the intensity-gradient vector $\vec{\nabla} I$, and $V^{\perp} = -\Delta I / |\nabla I|$. Once again, replacing ΔI by I_t leads to a similar equation involving the image-velocity $\vec{U}$.

Glazer also uses the Euler-Lagrange equations to transform this problem into a set of differential equations, and obtains a system of linear equations by using the finite-difference approach. The set of differential equations he obtains are,

$$\nabla I \left[(\nabla I)^T U + I_t \right] - \alpha^2 \begin{bmatrix} trace\{\nabla\nabla u\} \\ trace\{\nabla\nabla v\} \end{bmatrix} = 0, \tag{14}$$

where the operator $\nabla\nabla$ is as defined above.

Finally, Glazer uses a multi-resolution relaxation process to solve his system of equations, although his approach is more complex than Enkelmann's method and is based on recent theoretical work concerning general multi-level relaxation techniques. The hierarchical gradient-based approach and multi-level relaxation are both described in detail in [21].

5.2 Relating the gradient-based techniques to our framework

An important reason for developing a computational framework is the unification of a variety of different techniques so that we may be able to identify the elements that are common to these techniques and recognize their differences. In our particular case, the five

components described in section 2 are common to the various techniques that are unified within our framework; the techniques differ in the particular algorithmic choices made for the various components.

It should be evident from the description given in section 3 that the matching algorithm is completely consistent with the framework. In this section, we will show that both the hierarchical gradient-based techniques described above are also consistent with the framework; in the process of showing their consistency with the framework, we will also identify the algorithmic choices made in these two techniques for the various components. Finally, we will establish a mathematical relationship between the gradient-based techniques and the matching algorithm.

From the descriptions given above, it should be obvious that both Enkelmann and Glazer use spatial-frequency channels, and a coarse-to-fine control strategy; in particular, both use Gaussian low-pass filters, although the use of band-pass filters may be more appropriate, because they offer a greater separation of the spatial-frequencies. The control-strategy is similar to ours, except for some differences in the projection scheme. The use of the smoothness constraint is also explicit and the various formulations of the smoothness error functional are similar, the primary difference being the use of the weight matrix W by Enkelmann.

Although at first sight, the use of the term "match" criterion seems inappropriate for the gradient-based techniques, a careful examination of our definition of the term "match criterion" indicates that the use of the term is valid. According to our definition in section 2, a match criterion is the basis of computing local estimates of the displacements (and more generally, measurements of image motion). In both the gradient-based techniques under consideration, the local estimates are based on the *intensity constancy* assumption. However, in neither technique are the match criterion and the confidence measure explicit. Instead, these are implicitly present in the formulation of the minimization problems. Our purpose here is to explicitly identify these two components in each of the two gradient-based techniques, and demonstrate that an elegant mathematical relationship exists between our matching technique and the two gradient-based techniques.

Enkelmann's technique

As noted earlier, Enkelmann minimizes the sum of the two functionals E_{int} and E_{sm}, as defined in equations 7 and 8. Recall also that by using Euler-Lagrange equations, Enkelmann derives the differential equation shown in equation 9. It is easy to show (see

[7] for details) that solving this differential equation is also equivalent to minimizing $E' = E_{sm} + E'_{int}$, where E_{sm} is as defined by Enkelmann, and

$$E'_{int} = \iint dxdy(U - D)^T A(U - D). \tag{15}$$

In this definition, D is any vector such that $AD = -b$, and A and b are as defined in equations (10) and (11). By doing some further algebraic manipulation, it is easy to show that

$$E'_{int} = \iint dx\,dy \left[\lambda_1 \left((\vec{U} - \vec{D}) \cdot \hat{e}_1\right)^2 + \lambda_2 \left((\vec{U} - \vec{D}) \cdot \hat{e}_2\right)^2\right] \tag{16}$$

where λ_1 and λ_2 are the two eigenvalues of A, and $\hat{e}_1$ and $\hat{e}_2$ are the associated unit eigenvectors.

The transformation of Enkelmann's E_{int} to the form shown in equation 16 allows us to explicitly identify the match constraint and the confidence measures used in his technique. Specifically, the following interpretation is possible: the local estimates of displacement are the solutions to the equation $AU = -b$; corresponding to the components of the solution vector along the directions of the eigenvectors $\hat{e}_1$ and $\hat{e}_2$ of A, we can associate confidence measure λ_1 and λ_2 respectively. Note that while we can guarantee that there is at least one solution to this equation (see [7]), there is no guarantee that such a solution will be unique. In particular, it is easy to see that if both the eigenvalues of A are non-zero, there will be a unique solution vector (D), whereas if any of the eigenvalues are zero, the component of the solution vector along the direction of the corresponding vector can be arbitrarily chosen. In fact, the underlying image pixel can be regarded as a corner, edge, or a homogeneous area according to whether the matrix A has 2, 1, or 0 non-zero eigenvalues.

Glazer's technique

The definition of E_{int} used by Glazer is the following:

$$E_{int} = \iint dx\,dy\,|\nabla I|^2(U^\perp - V^\perp)^2, \tag{17}$$

where $V^\perp = -\Delta I/|\nabla I|$. If the parameter $\bar{x}^2$, which represents the window-size in the definitions of A and b is set to zero, the equation $AU = -b$ reduces to the equation

$$(\nabla I)^T U = -I_t,$$

which is equivalent to the normal-flow equation (6). Moreover, it can be shown that in this case,

$$\lambda_1 = 0,\ \lambda_2 = |\nabla I|^2, \text{ and } \hat{e}_1 = \hat{e}_{\nabla I}. \tag{18}$$

Thus, we see that Glazer's choice for the local estimates of motion and the confidence measures are similar to that of Enkelmann, with the important difference that the size of the window representing a point tends to zero, i.e., the window shrinks to a point.

The mathematical relationship between the matching and the gradient-based approaches

Thus far, we have shown that the two gradient-based techniques contain all the components of our framework, and are therefore completely consistent with it. In addition, we have also shown that Glazer's first-order gradient-based estimates of motion and the associated confidences are the values obtained by using Enkelmann's approach in the limiting case (when the window-size tends to zero). Here, we also note that there is a close mathematical relationship betwen our matching approach and the gradient-based approaches. The following theorem summarizes this relationship:

Theorem *In the limit, when the inter-frame time interval tends to zero, the formulation of the approximation error for image displacements used in the discrete matching approach converges to the second-order formulation of E_{int} for image-velocities used in the gradient-based approach, provided the third and higher order spatial intensity derivatives are ignored. Further, when the window-size represented by $\bar{x}^2$ tends to zero, E_{app} converges exactly to the first order gradient-based formulation of E_{int}.*

The proof of this theorem is too long for this paper, and can be found in [7]. The general approach is based on a derivation of Nagel [33] that in the limiting case (when the inter-frame time interval tends to zero), the minimization of the SSD measure is equivalent to solving the equation $AU = -b$.

5.3 Discussion

Thus far, we have attempted to establish the consistency of the gradient-baed techniques with our framework, and describe the mathematical relationship betwen the gradient-based techniques and our matching technique. In this section, we consider the significance of these results.

We have also shown that our framework provides a unifying perspective for the correlation matching techniques and gradient based techniques. In particular, the use of multi-resolution, multiple-frequency computations, the implicit or the explicit use of a confidence measure, and a smoothness constraint which uses that confidence measure seem essential for the success of all of these techniques.

We have also shown that in the gradient-based techniques, the confidence measure can be isolated from the smoothness assumption. This allows us to retain the confidence measure (which we believe to be essential), but reexamine the formulation of the smoothness assumption, and even consider whether such an assumption is always useful. For example, alternate forms of the smoothness constraint may be easily combined with the local measurements and the associated confidence measures: these include the flow-analyticity constraint of Waxman [43], the stochastic relaxation approaches of [19,30], and other such methods. Alternatively, we can postpone the construction of a dense displacement field, and use the local measurements and their confidences in a segmentation and grouping technique such as that of Adiv [2].

6 Processing Discontinuities in Image Motion

In this section, we consider one of the major unsolved problems in the analysis of visual motion and its relation to our computational framework. This situation involves processing discontinuities in image motion, which are present at the boundaries of surfaces, or at the boundaries of objects. Around the locations of such discontinuities, the smoothing involved in the spatial-frequency decomposition process creates intensity structures that have no physical correlates. Therefore, the information contained in the lower-frequency channels in the two images will be inconsistent. The local translational assumption and the spectral continuity principles are also violated. Obviously, it would be inappropriate to apply the smoothness constraint across such boundaries.

In order to process discontinuities in image motion, we must first detect such discontinuities. It should be clear from the brief discussion above that the detection of discontinuities cannot be postponed until after the computation of a dense flow field; rather, it should happen simultaneously with that computation. This means that our framework (as well as the techniques consistent with it) should be modified to incorporate the notion of discontinuities in image motion. Here we outline some possible ways to approach this task.

In the current version of our matching algorithm, the confidence measures are normal-

ized according to the magnitude of the minimum SSD value computed during the search process. The minimum SSD value is likely be large if (i) the search area does not contain the true match (either due to occlusion, or due to incorrect coarse-level estimate), (ii) the magnitudes of rotation and translation in depth are large, and (iii) the SNR (signal-to-noise) ratio is low. All these are cases where even the best match (and even if the SSD surface has a sharp minimum) is unreliable. In addition to the magnitude of the minimum SSD value, a significant difference in the shapes of the auto- and cross- SSD surfaces may also indicate an unreliable match.

A local measure of the likelihood of a false match is itself not sufficient to determine the presence of discontinuities. Typically the discontinuities will form a contour on the image plane. The spatial derivative of the image flow field across such a bounding contour can be expected to be large. If these three types of information are combined together, the robust detection of discontinuities may be possible.

One way to combine various sources of information to detect discontinuities in image motion is to first obtain a local "no match" measure, and explicitly include discontinuities in the smoothness process. This is somewhat similar to the approach suggested by Geman and Geman [19] for edge, and more recently by Marroquin *et al.* [30] for surface reconstruction from sparse depth data. The advantage of this approach is that by using two coupled models, one for the flow field, and the other for the discontinuity boundaries, the discontinuity detection process is dynamic and explicitly encoded.

7 Summary

We have described a hierarchical computational framework for the determination of dense displacement fields from a pair of images. We have also developed a matching algorithm consistent with our framework and demonstrated its performance on real images.

Our framework is sufficiently general in order to unify the gradient-based and the matching techniques. In particular, we have shown that in addition to our own technique, two successful gradient-based techniques are consistent with our framework. We have also established a clear mathematical relationship between the gradient-based techniques and our matching technique.

At present, the detection and processing of discontinuities in image motion still appears to be difficult. Our current approach (and for that matter almost all the current approaches for the measurement of motion) is also not capable of processing scenes contain-

ing transparent or fence like surfaces. Finally, we have not addressed the issues involved in processing multiple-frames. While these problems form the obvious basis for further research in the measurement of visual motion, we wish to note that our algorithm is readily applicable for a large class of commonly encountered images and performs robustly (see [7]) with almost no adjustment of parameters.

ACKNOWLEDGEMENTS

The author wishes to thank Profs. Edward Riseman and Allen Hanson for their continued advice and support through the course of this work. Thanks are also due to Dr. George Reynolds, Prof. Riseman, Prof. Al Hanson and Michael Boldt for their comments on earlier drafts, to Dr. Richard Weiss for his help with some of the mathematical aspects of this research, and Dr. Mark Snyder for clarifying some fundamental concepts in linear algebra. Finally, thanks are due to the members of the UMass VISIONS group for creating a unique and valuable research environment.

References

[1] Adelson E. H. and Bergen J. R. Spatiotemproal energy models for the perception of motion, *J. Opt. Soc. Am. A* Vol. 2, No. 2, pp. 284-299, 1985.

[2] Adiv G, Determining 3-d motion and structure from optical flows generated by several moving objects, *IEEE T-PAMI*, 7 (4), pp. 384-401, 1985.

[3] Adiv G., Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field, *Proc. CVPR*, pp. 70-77, 1985.

[4] Aggarwal J. K., Davis L. S., and Martin W. N., Correspondence processes in dynamic scene analysis, *Proc. IEEE*, Vol. 69, No. 5, pp. 562-572, 1981.

[5] Anandan P., Computing dense displacement fields with confidence measures in scenes containing occlusion, *SPIE Intelligent Robots and Computer Vision Conference*, Vol. 521, pp 184-194, 1984, also *COINS Technical Report 84-32*, University of Massachusetts, December 1984.

[6] Anandan P. and Weiss R., Introducing a smoothness constraint in a matching approach for the computation of displacement fields, *DARPA IU Workshop Proc.*, pp. 186-196, 1985.

[7] Anandan P., Measuring visual motion from image sequences, *Ph.D. dissertation*, COINS Department, University of Massachusetts, Amherst, Mass., *in preparation.*

[8] 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.

[9] Beaudet, P. Rotationally invariant image operators, *Proc. International Conference on Pattern Recognition*, 1978, pp. 579-583.

[10] Boldt M., Token based image abstraction, *COINS Tech. Report*, University of Massachusetts, *in preparation.*

[11] Burt P. J., Yen C. and Xu X., Local correlation measures for motion analysis: A comparative study, *IEEE Proc. PRIP*, 269-274, 1982.

[12] Burt P. J., Hong T. H and Rosenfeld A., Image segmentation and region property computation by cooperative hierarchical computation, *IEEE Trans. Systems, Man, Cybernetics* 11, 1981, pp. 802-809.

[13] Burt P. J., Fast filter transforms for image processing, *Computer graphics and image processing*, 16, pp. 20-51, 1981.

[14] Burt P. J., Yen C. and Xu X., Multi-resolution flow-through motion analysis, *IEEE CVPR Conference Proceedings*, June 1983, pp. 246-252.

[15] Carmo M. P., *Differential geometry of curves and surfaces*, Prentice-Hall, New Jersey, 1976.

[16] Crowley J. L., and Stern R. M., Fast computations of the difference of low-pass transform, *IEEE Transactions on PAMI*, vol. PAMI-6, pp. 212-222, 1984.

[17] Enkelmann, W., Investigations of multigrid algorithms for the estimation of optical flow fields in image sequences, *Proc. of the Workshop on Motion: Representation and Control*, S. Carolina, pp. 81-87, 1986.

[18] Fang J. and Huang T. S., Some experiments on estimating the 3-d motion parameters of a rigid body from two consecutive Image Frames, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, NO. 5, pp 545-554, September, 1984.

[19] Geman S. and Geman D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721-741, 1984.

[20] Glazer F., Reynolds G. and Anandan P., Scene matching by hierarchical correlation, *IEEE CVPR conference*, June 1983, pp. 432-441.

[21] Glazer F., Hierarchical motion detection, *Ph.D. dissertation*, COINS Department, University of Massachusetts, Amherst, Ma., February 1987.

[22] Grimson W. E. L., Computational experiments with a feature based stereo algorithm, *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, No. 1, pp. 17-34, 1985.

[23] Hanson A.. R and Riseman E. M., Processing cones: A computational struture for image analysis, in: *Structured computer vision* Tanimoto S. and Klinger A. (Eds.), Academic Press, New York, 1980.

[24] Heeger D., and Pentalnd A., Seeing structure through chaos, *Proc. of the Workshop on motion: Representation and Control*, S. Carolina, pp. 131-136, 1986.

[25] Hildreth E. C., The measurement of visual motion, *Ph.D. dissertation*, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, Ma., 1983.

[26] Horn B. K. P., and Schunck B. G., Determining Optical Flow, *Artificial Intelligence*, vol. 17, pp. 185-203, 1981.

[27] Klinger A., and Dyer R. D., Experiments on picture respresentations using regular decomposition, *Computer graphics and image processing*, vol 5, no. 1, pp. 68-105, 1976.

[28] Lucas B. D., and Kanade T., An iterative image registration technique with an application to stereo vision, *Proc. 7th IJCAI*, Vancouver, B. C., Canada, pp. 674-679, 1981.

[29] Marr D. and Poggio T., A computational theory of human stereo vision, *Proc. Roy. Soc. London, Ser. B*, 204, pp 301-308, 1979.

[30] Marroquin J., Mitter S., and Poggio T.,, Probabilistic solution for ill-posed problems in computational vision, *Proc. DARPA IU Workshop*, Miami Beach, Fl., pp. 293-309, 1986.

[31] Mayhew J. E. W. and Frisby J. P., Psychophysical and computational studies towards a theory of human stereopsis, *Artificial Intelligence*, Vol. 17, pp. 349-385, 1981.

[32] Moravec H. *Robot rover visual navigation*, UMI Research press, Ann Arbor, Michigan, 1981.

[33] Nagel H. H., Displacement vectors derived from second order intensity variations in image sequences, *CVGIP*, vol. 21, pp. 85-117, 1983.

[34] Nagel H. H. and Enkelmann W., An investigation of smoothness constraints for the estimation of displacement vector fields from image sequnces, *IEEE transactions on PAMI*, vol. PAMI-8, pp. 565-593, 1986.

[35] Nagel H. H., Image sequences – Ten (Octal) years – from phenomenology towards a theoretical foundation, *Proc. of Eighth ICPR*, Paris, France, 1986.

[36] Rosenfeld A. and Kak A. C., *Digital picture processing*, Academic Press, New York, 1976.

[37] Tanimato S. L. and Pavlidis T., A hierarchical data structure for picture processing, *CGIP*, vol. 4, no. 2, pp. 104-119, 1975.

[38] Terzopoulos D., Mutiresolution Computation of Visible-Surface Representations, *Phd Dissertation*, Massachusetts Institute of Technology, Jan. 1984.

[39] Ullman S., Analysis of visual motion by biological and computer systems, *IEEE computer*, pp. 57-69, 1981.

[40] van Santen J. P. H. and Sperling G., Elaborated Reichhardt detectors, *J. of Opt. Soc. Am.*, vol. 2, no. 7, pp. 300-321, 1985.

[41] Watson A.B. and Ahmuda A. J., Model of human visual-motion sensing, *J. of Opt. Soc. Am.*, vol. 2, no. 7, 1985.

[42] Waxman A., An image-flow paradigm, *Proc. Workshop on computer vision*, Annapolis, MD, pp. 49-57, 1984.

[43] A. Waxman and K. Wohn, Contour evaluation, neighborhood deformation, and global image flow: planar surfaces in motion, *CS-TR-1394*, University of Maryland, April 1984.

[44] Weiss R. and Boldt M., Geometric grouping applied to straight lines, *Proc. IEEE CVPR*, Miami Beach, Florida, pp. 489-493, 1986.

[45] Williams L. R. and Anandan P., A coarse-to-fine control strategy for stereo and motion on a mesh-connected computer, *IEEE Computer Vision and Pattern Recognition Conference*, Miami Beach, Florida, pp. 219-226., 1986.

[46] Wong R. Y. and Hall E. L., Sequential hierarchical scene matching, *IEEE transactions on Computers*, Vol. 27, No. 4, pp. 359-366, 1978.