Combining Stereo and Monocular Information to Compute Dense Depth Maps that Preserve Depth Discontinuities

Pascal Fua (fua@mirsa.inria.fr)*
INRIA Sophia-Antipolis
2004 Route des Lucioles
06565 Valbonne Cedex
France

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
USA

Abstract

In this paper, we show how simple and parallel techniques can be efficiently combined to compute dense depth maps and preserve depth discontinuities in complex real world scenes.
Our algorithm relies on correlation followed by interpolation. During the correlation phase the two images play a symmetric role and we use a validity criterion for the matches that eliminates gross errors: at places where the images cannot be correlated reliably, due to lack of texture or occlusions for example, the algorithm does not produce wrong matches but a very sparse disparity map as opposed to a dense one when the correlation is successful. To generate dense depth map, the information is then propagated across the featureless areas but not across discontinuities by an interpolation scheme that takes image grey levels into account to preserve image features.
We show that our algorithm performs very well on difficult images such as faces and cluttered ground level scenes. Because all the techniques described here are parallel and very regular they could be implemented in hardware and lead to extremely fast stereo systems.

1 Introduction

Over the years numerous algorithms for passive stereo have been proposed, they can roughly be classified in two main categories [Barnard et al. 82]:

1. Feature Based. Those algorithms extract features of interest from the images, such as edge segments or contours, and match them in two or more views. These methods are fast because only a small subset of the image pixels are used, but may fail if the chosen primitives cannot be reliably found in the images; furthermore they usually only yield very sparse depth maps.
2. Area Based. In these approaches, the system attempts to correlate the grey levels of image patches in the views being considered, assuming that they present some similarity. The resulting depth map can then be interpolated. The underlying assumption appears to be a valid one for relatively textured areas; however if may prove wrong at occlusion boundaries and within featureless regions.

   Alternatively the map can be computed by directly fitting a smooth surface that accounts for the disparities between the two images. This is a more principled approach since the problem can be phrased as an optimization one; however the smoothness assumptions that are required may not always be satisfied.

All these techniques have their strengths and weaknesses and it is difficult to assess their compared merits since few researchers work on similar data sets. However, one can get a feel for the relative performances of these systems from the study by Guelch [Guelch 88]. In this work, the author has assembled a standardized data set and sent it to 15 research institutes across the world. It appears that the correlation based system developed at SRI by Hannah [Hannah 88] has produced the best results both in terms of precision and reliability. Unfortunately this system only matches a very small proportion, typically less than 1%, of the image points.

In this paper we propose a correlation algorithm that reliably produces far denser maps with very few false matches and can therefore be effectively interpolated. In the next section we describe our hypothesis generation mechanism that attempts to match every point in the image and uses a consistency criterion to reject invalid matches. This criterion is designed so that when the cor-

*This research was supported in part under the Centre National d'Etudes Spatiales VAP contract and in part under a Defense Advanced Research Projects Agency contract.

relation fails, instead of yielding an incorrect answer, the algorithm returns NO answer. As a result, the density of the computed disparity map is a very good measure of its reliability. The interpolation technique described in the section that follows combines the depth map produced by correlation and the grey level information present in the image itself to introduce depth discontinuities and fit a piecewise smooth surface. These algorithms have proven very effective on real data. Their parallel implementation on a Connection Machine[tm1] relies only on local operations and on nearest neighbor communication; they could be ported to a dedicated architecture, thereby making fast and cheap systems possible.

2 Correlation

Most correlation based algorithms attempt to find interest points on which to perform the correlation. While this approach is justified when only limited computing resources are available, with modern hardware architectures and massively parallel computers it becomes possible to perform the correlation over the whole image and retain only results that appear to be "valid." The hard problem is then to provide an effective definition of what we call validity and we will propose one below.

In our approach, we compute correlation scores for every point in the image by taking a fixed window in the first image and a shifting window in the second. The second window is moved in the second image by integer increments along the epipolar line and an array of correlation scores is generated. In this work we use correlation of grey level values and take the correlation score to be $s = \max(0, 1 - c)$ where

$$c = \frac{\overline{((I_1 - \overline{I_1}) - (I_2 - \overline{I_2}))^2}}{\sqrt{\overline{(I_1 - \overline{I_1})^2}\,\overline{(I_2 - \overline{I_2})^2}}} \quad , \qquad (1)$$

I_1 and I_2 being the left and right image intensities, X the average value of X over the correlation window and dx,dy the displacement along the epipolar line.[2] The measured disparity can then be taken to be the one that provides the highest value of s. In fact, to compute the disparity with subpixel accuracy, we fit a second degree curve to the correlation scores in the neighborhood of the maximum and compute the optimal disparity by interpolation.

2.1 Validity of the Disparity Measure

As shown by Nishihara [Nishiharaet al. 83], the probability of a mismatch goes down as the size of the correlation window and the amount of texture increase. However, using large windows leads to a loss of accuracy and the possible loss of important image features. For smaller windows, the simplest definition of validity would call for a threshold on the correlation score; unfortunately such a threshold would be rather arbitrary and, in practice, hard to choose. Another approach is to build a correlation surface by computing disparity scores for a point in the neighborhood of a prospective match and checking that the surface is peaked enough [Anandan 89]. It is more robust but also involves a set of relatively arbitrary thresholds. Here we propose a definition of a valid disparity measure in which the two images play a symmetric role and that allows us to reliably use small windows. We perform the correlation twice by reversing the roles of the two images and consider as valid only those matches for which we measure the same depth at corresponding points when matching from I_1 into I_2 and I_2 into I_1. As shown in Figure 1, this can be defined as follows.

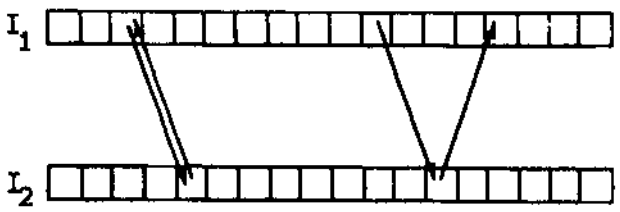


Figure 1: Consistent vs inconsistent matches: the two rows represent pixels along two epipolar lines of I_1 and I_2 and the arrows go from a point in one of the images towards the point in the other image that maximizes the correlation score. The match on the left is consistent because correlating from /\ to I2 and from I_2 to /\ yields the same match unlike the matches on the right that are inconsistent.

> Given a point P\ in I_1, let P_2 be the point of I_2 located on the epipolar line corresponding to P\ such that the windows centered on P\ and P_2 yield the optimal correlation measure. The match is valid if and only if P\ is also the point that maximizes the score when correlating the window centered on P_2 with windows that shift along the epipolar line of I_1 corresponding to P_2

For example, the validity test is likely to fail in presence of an occlusion. Let us assume that a portion of a scene is visible in I_1 but not I_2. The pixels in /\ corresponding to the occluded area in I_2 will be matched, more or less at random, to points of I_2 that correspond to different points of I_1 and are likely to be matched with them. The matches for the occluded points will therefore be declared invalid and rejected. We illustrate this behaviour using the portion of the tree scene of Figure 2 outlined in Figure 2(a). Different parts of the ground be-

[1] TMC Inc.

[2] We remove the mean to offset transformations of the images that may result from slightly different settings of the cameras.

(a) (b) (c)

Figure 2: (a) An outdoor scene with two trees and a stump. The same scene seen from the left (b) and the right (c) so that different parts of the ground are occluded by the trees.

tween the two trees and between the trees and the stump are occluded in Figures 2 (b) and (c). In Figures 3 (a) and (b), we show the computed disparities for this image window after correlation with the images shown in Figures 2(b) and 2(c) respectively. The points for which no valid match can be found appear in white and the areas where their density becomes very high correspond very closely to the occluded areas for both pairs of images. These results have been obtained using 3x3 correlation windows; these small windows are sufficient in this case because the scene is very textured and gives our validity test enough discriminating power to avoid errors.

We use the face shown in Figure 4 to demonstrate another case in which the validity test rejects false matches. The epipolar lines are horizontal and in Figure 4(d) we show the resulting disparity image, using 7x7 windows, in which the invalid matches appear in black. In Figure 4(e) we show another disparity image computed after having shifted one of the images vertically by two pixels, thereby degrading the calibration and the correlation. Note that the disparity map becomes much sparser but that no gross errors are introduced. In practice, we take advantage of this behaviour for poorly calibrated images: we compute several disparity maps by shifting one of the images up or down and retaining the same epip>olar lines,[3] thereby replacing the line by a band, and retain the highest scoring valid matches.[4]

In the two examples described above, we have shown that when the correlation between the two images of a stereo pair is degraded our algorithm tends, instead of making mistakes, to yield sparse maps. Generally speaking, correlation based algorithms rely on the fact that the same texture can be found at corresponding points in the two images of a stereo pair and are known to fail when:

- The areas to be correlated have little texture.
- The disparities vary rapidly within the correlation window.
- There is an occlusion.

If we consider the local image texture as a signal to be found in both images, we can model these problems as noise that corrupts the signal. In a companion report [Fua 9l], we use synthetic data to formalize this argument and show that as the noise to signal ratio increases, or equivalently, as the problems mentioned above become more acute, the performance of our correlation algorithm degrades gracefully in the following sense:

> As the signal is being degraded, the density of matches decreases accordingly but the ratio of correct to false matches remains high until this proportion has dropped very low.

In other words, a relatively dense disparity map is a *guarantee* that the matches are correct, at least up to the precision allowed by the resolution being used. In the report [Fua 9l], we also show the effectiveness of a very simple heuristic: if we reject not only invalid matches but also isolated valid matches we can increase even more the ratio correct/incorrect matches without losing a large number of the correct answers.

Other stereo systems include a validity criterion similar to ours but use it as only one among many others. In our case, because we correlate over the whole image and not only at interest or contour points, we do not need the other criteria and can rely on density alone. However, our validity test depends on the fact that it is

[3] assumed not to be exactly vertical

[4] The precision of the computed distance is then obviously degraded, but remains qualitatively correct for large enough baselines.

Figure 3: (a) The result of matching 2(a) and 2(b) for window of 2(a) delimited by the white rectangle, (b) The result of matching 2(a) and 2(c) for the same windows, (c) The merger of four disparity maps computed using the image of Figure 2 (a) as a reference frame, the two other images of 2 and two additional images. Invalid matches appear in white and become almost dense in occluded areas of (a) and (b). The closest areas are darker; note that they are few false matches although the correlation windows used in this case are very small (3x3).

improbable to make the same mistake twice when correlating in both directions and can potentially be fooled by repetitive patterns, which is a problem we have not addressed yet.

2.2 Hierarchical Approach and Additional Images

To increase the density of our potentially sparse disparity map, we use windows of a fixed size to perform the matching at several levels of resolution,[5] which is almost equivalent to matching at one level of resolution with windows of different sizes as suggested by Kanade [Kanade et al. 90] but computationally more efficient. More precisely, as shown by Burt [Burt et al. 82], it amounts to performing the correlation using several frequency bands of the image signal.

We then merge the disparity maps by selecting, for every pixel, the highest level of resolution for which a valid disparity has been found. In Figure 4 (c) we show the merger of the disparity maps for two levels of resolution that is dense and exhibits more of the fine details of the face than the map of figure 4 (e) computed using only the coarsest level of resolution. The reliability of our validity test allows us to deal very simply with several resolutions without having to introduce, as in [Kanade et al. 90] for example, a correction factor accounting for the fact that correlation scores for large windows tend to be inferior to those for small windows.

The computation proceeds independently at all levels of resolution and this is a departure from traditional hierarchical implementations that make use of the results generated at low resolution to guide the search at higher resolutions. While this is a good method to reduce computation time, it assumes that the results generated at low resolution are more reliable, if less precise, than those generated at high resolution; this is a questionable assumption especially in the presence of occlusions. For example in the case of the trees of Figure 2, it could lead to a computed distance for the area between the trunks that would be approximately the same as that of the trunks themselves, which would be wrong. Furthermore, it can be shown [Fua 91] that, in the absence of repetitive patterns, the output of our algorithm is not appreciably degraded by using the large disparity ranges that our approach requires.

As suggested by several researchers, more than two images can and should be used whenever practical. When dealing with three images or more, we take the first one to be our reference frame, compute disparity maps for all pairs formed by this image and one of the others and merge these maps in the same way as those computed at different levels of resolution. In this way we can generate a dense disparity map, such as the one of Figure 3 (c): the three images of Figure 2 belong to a series of five taken by an horizontally moving camera. Taking the image of 2(a) as our reference frame, we merge the four resulting disparity maps, each of them relatively sparse, to produce a dense map with few errors.

In this section we have presented an hypothesis generation mechanism that produces depth maps that are correct where they are dense and unreliable only where they become very sparse. Typically these sparse measurements occur in featureless areas that are usually smooth and at occlusion boundaries where one expects to find an image intensity edge. To fit a surface, one must therefore

[5]Computed by subsampling gaussian smoothed images.

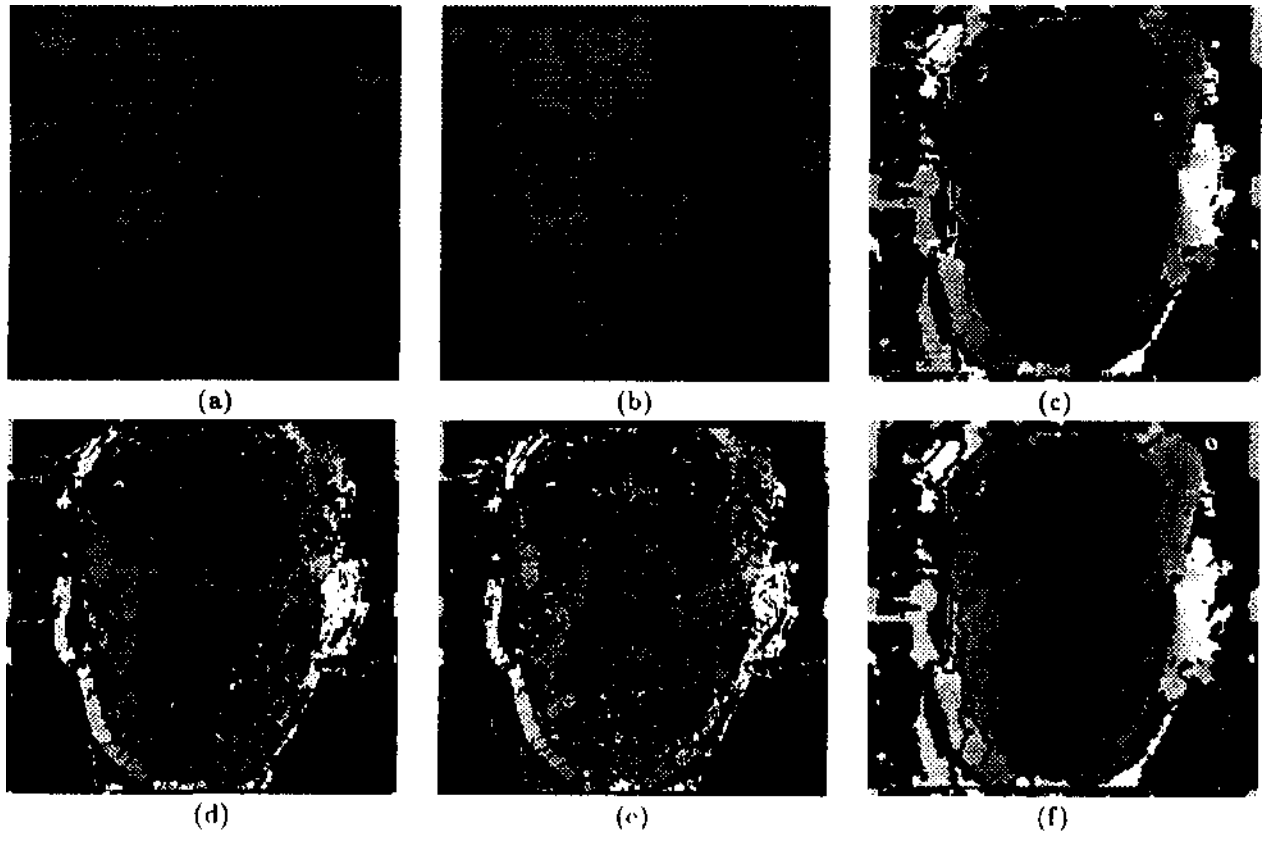


Figure 4: (a) (b) Left and right 256x256 images of a face, (c) The disparity map obtained by merging the results computed at two levels of resolution, (d) Disparity map computed at the highest resolution, (e) Disparity map computed at the highest resolution after shifting the right up by one pixel, (f) Disparity map computed at the lower resolution.

interpolate those measures in such a way as to propagate the depth information in the featureless areas and preserve depth discontinuities. In the next section, we describe the model and algorithm we use to perform the interpolation.

3 Interpolation

We model the world as made of smooth surfaces separated by depth discontinuities. We also assume that these depth discontinuities produce gradients in grey level intensities due to changes in orientation and surface material. We first describe a simple interpolation model that is well suited to images with sharp contrasts and then propose a refinement of that scheme for lower contrast scenes.

3.1 Simple Interpolation Model

Ideally, if we could measure with absolute reliability the depth, w0, at a number of locations in the image, we could compute a depth image *w* by minimizing the following criterion:

$$
\begin{aligned}
C &= \int s(w - w0)^2 + \lambda_x(\frac{\partial w}{\partial x})^2 + \lambda_y(\frac{\partial w}{\partial y})^2 \qquad (2)\\
s &= 1 \text{ if } w0 \text{ has been measured}, 0 \text{ otherwise.}\\
\lambda_x &= 0 \text{ if horizontal discontinuity}, c_x \text{ otherwise.}\\
\lambda_y &= 0 \text{ if vertical discontinuity}, c_y \text{ otherwise.}
\end{aligned}
$$

where c_x and c_y are two real numbers that control the amount of smoothing.

As discussed in the previous section, when a valid disparity can be found it is reliable and can be used, along with the camera models, to estimate $w0$; we then take .s to be the normalized correlation score of Equation 1. Assuming that changes in reflectance can be found at depth discontinuities, we replace the λ_x and λ_y of Equation 2, by terms that are inversely proportional to the image gradients in the *x* and *y* directions. In fact, we have observed that the absolute magnitudes of the gradients are not as relevant to our analysis as their local relative magnitudes: boundaries can be adequately characterized as

the locus of the strongest local gradients, independent of the actual value of these gradients. We therefore write:

$$\lambda_x = c_x F_{Norm}(\frac{\partial I}{\partial x}) \qquad (3)$$
$$\lambda_y = c_y F_{Norm}(\frac{\partial I}{\partial y})$$

where F_{Norm} is the piecewise linear function defined by

$$F_{Norm}(x) = \begin{cases} 1 & \text{if} \quad x < x_0 \\ \frac{x_1 - x}{x_1 - x_0} & \text{if} \quad x_0 < x < x_1 \\ 0 & \text{if} \quad x_1 < x \end{cases} \qquad (4)$$

and x_0 and x_1 are two constants. In all our examples, x_0 IS the median value of x in the image and x_1 its maximum value . The result is quite insensitive to the value chosen for x_0 as long as it does not become so large as to force the algorithm to ignore all edges. What really matters is the monotoniaty of F_{Norm} that allows the depth information to propagate faster in the directions of least image gradient and gives to the algorithm a behaviour somewhat similar to that of adaptative diffusion schemes (e.g [Perona et al. 87]).

To compute *W,* the vector of all values of w, we discretize the quadratic criterion *C* of Equation 2; we then solve, using a conjugate gradient method [Szeliski 90; Terzopoulos 86], the linear equation

$$\frac{\partial C}{\partial W} = 0 \qquad (70$$

In Figure 3.1, we show the depth map computed by interpolating the disparity map of Figure 4(c). Note that the main features of the face, nose, eyebrows and mouth have been correctly recovered.

This simple interpolation technique is appropriate for the face of Figure 4 that presents few low-contrast depth discontinuities but produces a somewhat blurry result for the tree scene of Figure 2, as can be seen in Figure 6(a) To improve upon this situation, we propose a slightly more elaborate interpolation scheme that takes depth discontinuities explicitly into account.

3.2 Introducing Depth Discontinuities

The λ_x and λ_y coefficients defined by Equation 3 introduce "soft" discontinuities: when the contrast is low, some smoothing occurs across the discontinuity. The depth image, however, is less smoothed than in the complete absence of an edge resulting in a strong w gradient. We take advantage of this property of our "adaptative" smoothing by defining the following iterative scheme:

1. Interpolate using the λ_x and λ_y defined above.
2. Iterate the following procedure:
   (a) Recompute λ_x and λ_y as functions of both the intensity gradient and the depth gradient of the interpolated image:

$$\lambda_x = F_{Norm}(\frac{\partial I}{\partial x}) F_{Norm}(\frac{\partial w}{\partial x})^{\alpha} \qquad (6)$$
$$\lambda_y = F_{Norm}(\frac{\partial I}{\partial y}) F_{Norm}(\frac{\partial w}{\partial y})^{\alpha}$$

   where α is a constant equal to 2 in our examples.

   (b) Interpolate again the raw disparity map using the new λ_x and λ_y coefficients

The algorithm converges after a few iterations resulting in a much sharper depth map such as the one of Figure 6(h). This algorithm can be regarded as a continuation method on the depth discontinuities. We start without knowing their location, use the grey level information to hypothesize them and then propagate the results.

4 Conclusion

In this work we have described a correlation based algorithm that combines two simple and parallel techniques to yield reliable depth maps in the presence of depth discontinuities, occlusions and featureless areas:

- The correlation is performed twice over the two images by reversing their roles and only matches that are consistent in both directions are retained, thereby guaranteeing a very low error rate
- The disparity map is then interpolated using a technique that, takes advantage of the grey level information present in the image to preserves depth discontinuities and propagate the information across featureless areas.

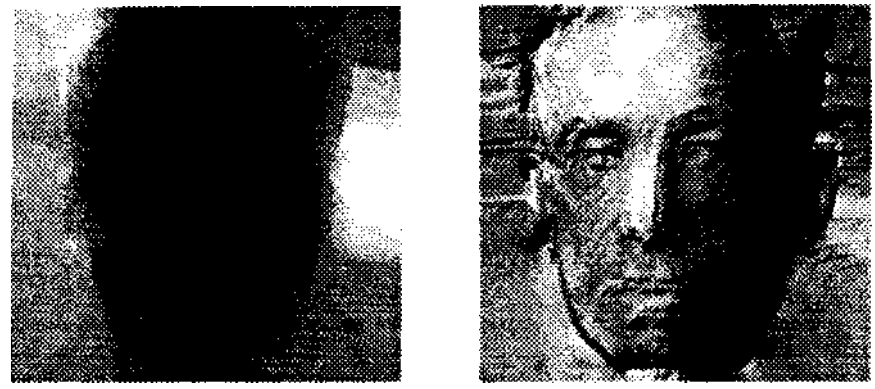
The depth maps that we compute are qualitatively correct and the density of acceptable matches provides us with an excellent estimate of their reliability. Because of the great regularity and simplicity of the techniques described here,[6] we hope to be able to build dedicated hardware that would implement them and could, for example, be used by a mobile robot in an outdoor environment.

References

[Anandan 89] P. Anandan. A computational framework and an algorithm for the measurement of motion. *International Journal of Computer Vision,* 2(3):283 310, 1989.

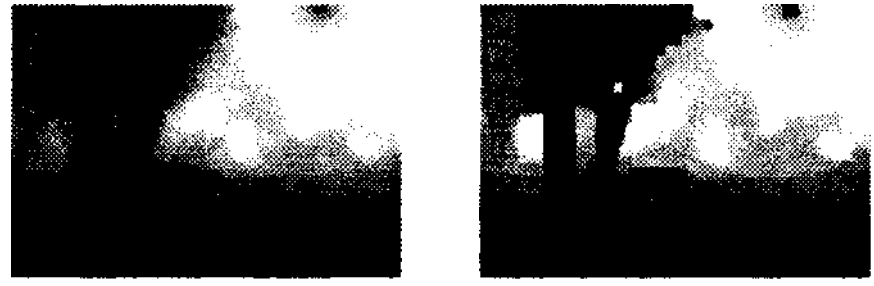
[Barnard et al. 82] ST. Barnard and MA Fischler. Computational stereo. *Computational Surveys, 14*(4):553-572, 1982.

[6]All the algorithms described here are implemented in *lisp on a Connection Machine and use exclusively local operations and nearest neighbor communication.



(a) (b)

Figure 5: (a) Interpolated depth map for the face of Figure 4 (b) Shaded views generated by shining a light on the surface.



(a) (b)

Figure 6: (a) Trees depth image computed by smoothing using the algorithm of section 3.1. (b) Depth image after four iterations of the iterative scheme of section 3.2. We have stretched the depth images to enhance the contrast so that the furthest areas appear completely white. Note that the trunks and the stump clearly stand out.

[Burt et al. 82] P.J. Burt, C. Yen, and X. Xu. Local correlation measures for motion analysis. In *IEEE PRIP Conference,* pages 269-274, 1982.

[Fua 91] P. Fua. *A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features.* Research Report 1369, INRIA, January 1991.

[Guelch 88] E. Guelch. Results of test on image matching of isprs wg iii / 4. *International Archives of Photogrammetry and remote sensing,* 27(III):254-271, 1988.

[Hannah 88] M.J. Hannah. Digital stereo image matching techniques. *International Archives of Photogrammetry and remote sensing,* 27(III).280-293, 1988.

[Kanade et al. 90] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptative window: theory and experiment. In *Image Understanding Workshop,* September 1990.

[Nishihara et al. 83] H.K. Nishihara and T.Poggio. Stereo vision for robotics. In *1SRR83 Conference,* Bretton Woods, New Hampshire, 1983.

[Perona et al. 87] *P.* Perona and J. Malik, Scale space and edge detection using anisotropic diffusion. In *IEEE Computer Society Workshop on Computer Vision,* pages 16-22, Miami, Florida, 1987.

[Szeliski 90] R. Szeliski. Fast surface interpolation using hierarchical basis functions. *IEEE IYansactions on Pattern Analysis and Machine Intelligence,* 12(6):513-528, June 1990.

[Terzopoulos 86] D. Terzopoulos. Image analysis using multigrid relaxation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 8(2):129- 139, March 1986.