Chapter 4

Accuracy: Dealing with Ambiguous Matches

Time flies like an arrow; fruit flies like a banana.
— Groucho Marx

A problem shared by virtually all stereo methods is that of *ambiguous matches,* or false targets. When two or more parts of an image pair are similar in appearance, as can happen when a repetitive pattern like a checkerboard or picket fence is present, a part of the pattern in one image might seem to match several parts in the other. When this happens, when there are *multiple* potential correspondents for a given pixel, an *ambiguous match* is said to exist. This problem arises because image points are compared in a purely local fashion, and the effects can be seen in disparity estimates that are wildly inaccurate.

In this chapter we will show examples of ambiguous matches, explain how their presence affects several stereo methods, and demonstrate that our phase-based method compensates for the problem more effectively than other methods. The *Disparity Space* provides the framework for this analysis, and motivates a generalized representation of disparity.

4.1 The Problem of Ambiguity

Ambiguous matches are second only to modeling and calibration errors as a source of inaccuracy in stereo vision systems. By modeling errors, we mean the violation of any assumptions about the image formation process that went into producing a particular stereo pair. Stereo methods by their nature make many implicit assumptions about their input images, e.g., that the units of horizontal and vertical spacing in the real world are identical, the left image was taken by the left camera, surfaces are diffuse, the lens caps have been removed, etc. And by calibration, we mean any computation that will influence (independently of the stereo image data) either the range of disparities being checked or the process by which they are checked in the stereo method. It should be obvious that any part of the image acquisition process that violates modeling assumptions or reduces the accuracy or precision of the calibration results could cause inaccurate results to be generated. Yet even when all of these constraints on the physical system have been met, the correspondence problem remains.

The correspondence problem lies at the heart of all stereo methods. In many methods the determination of correspondence is based solely upon information found in the stereo image pair, without knowledge of the actual 3D scene structure or image segmentation. Yet even though a stereo method returns the "best" answer according to its model, sometimes the "best" correspondence is not the correct one. When an improper correspondence is established, the resulting disparity can be *very* inaccurate.

This problem can be alleviated somewhat by incorporating more data and using better models in the stereo method. There are many such techniques in the literature: using many more than two cameras as in multibaseline stereo (Kanade et al., 1995; Webb, 1993), using more realistic models of image formation (Bhat & Nayar, 1995; Belhumeur, 1995, and Chapter 5 in this text), and interpreting the same data at several scales (the subject of the following sections). Although they differ in their approaches, all of these techniques share the common goal of finding the single best disparity estimate for a given pixel.

But sometimes the "best" answer is not the desired one. Figure 4.1 shows an artificial example of this phenomenon. In the figure a stereo image pair is actually embedded into a single image; this type of image is called an *autostereogram*.[1] In an autostereogram, it would appear that the left and right stereo images are identical. And when stereo processors are given the same left and right stereo images, most (including your eyes) will conclude that the "best" disparity map is a uniform zero shift, i.e., the images contain no interesting 3D structure. However, in this case the interesting stuff happens when you realize that the left and right images of interest are *not* identical; in Figure 4.1, the left image does indeed start

[1] Some people have difficulty seeing the depth in images like this. Give it your best shot.

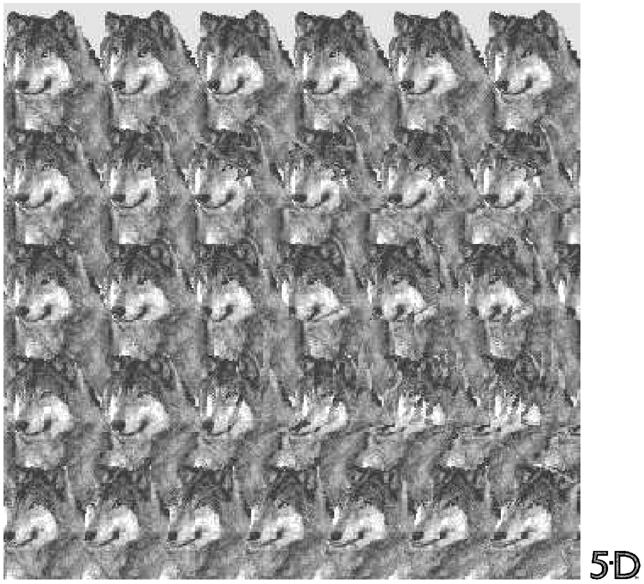


Figure 4.1: A 340x340 autostereogram of a wolf. See Figure 4.24 for an elucidation of the embedded structure. Reprinted with permission from Blue Mountain Arts, Inc.

at the left border, but the right image starts about one sixth of the way in. Running a coarse to fine stereo method on such pre-shifted images brings out the structure that was embedded into the original figure (see Figure 4.24 after trying it yourself), and demonstrates that in this case it is not the "best" disparity that is of interest, but rather the "second best."

The rest of this section will give us the language and concepts required for our discussion of ambiguity. The sections that follow will discuss both the impact of ambiguity on stereo methods, and extensions to stereo methods that will model, if not eliminate, problems due to ambiguous matches.

4.1.1 Definitions

An *ambiguous match* occurs when a stereo method is unable to determine a unique correspondent for a particular pixel. Ambiguity can be viewed in two ways: as an inherent property of an image, or as an artifact of a particular stereo method. It is certainly possible to construct inherently ambiguous stereo images. Consider two images of a frontoplanar checkerboard, where the checkerboard occupies the entire fields of view: the alignment of the squares cannot be determined from the images alone. Naturally, these types of images will cause problems for stereo matchers, but they are not the only source of ambiguity.

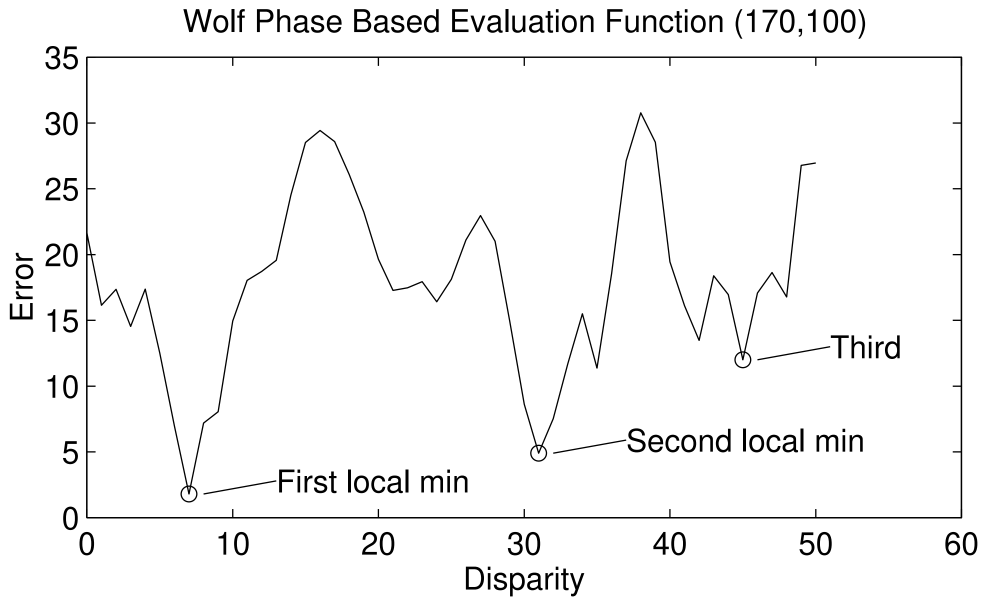


Figure 4.2: Evaluation Function for pixel (170,100) of Figure 4.1 (with its pre-shifted counterpart), computed using our phase-based method with linear frequency samples. This information is presented in the context of a complete scanline in column 100 of Figure 4.3. Darker pixels denote less error, so the dark stripe from pixel number 75 to 200 at disparity 7 represents the best guess.

The limitations of stereo matchers can also cause image regions in stereo imagery to appear ambiguous.

Informally, ambiguous matches will occur when the small area around one pixel "looks like" more than one area in the second image. This is because most stereo algorithms reduce the computational complexity of the matching process by determining correspondence using only small image patches at some stage of their processing. Even the so-called multi-scale coarse to fine methods only make local decisions, based on one or possibly two levels of hierarchy. It is this narrowing of attention, the making of a hard decision based on local information at a single scale, that makes stereo algorithms especially susceptible to ambiguous matches.

The chance of encountering ambiguity looms large in images with repetitive texture, since many image patches will appear to be nearly identical, but ambiguity can also occur even when the images do not exhibit obvious repetition. For many algorithms, all that is necessary is that two small patches appear identical, where "small" and "identical" are defined by the search window size and evaluation function used by the algorithm.

The presence of ambiguous matches can be determined by inspecting a pixel's disparity space, i.e., the result of the evaluation function. If the evaluation function profile is unimodal, i.e., if it has a single minimum and a monotonically increasing first derivative, then a unique disparity exists at that minimum and there is no ambiguity. However, if as is often the case the evaluation function profile is *not* unimodal, then any of the alternate local minima must be considered potential matches as well, and therefore an ambiguous match exists (see Figure 4.2). The degree to which the match at that pixel is ambiguous can be informally determined by considering the ratio of values of the evaluation function at the lowest minima: the closer the ratio is to 1, the more ambiguous is the match.

$$\textit{Ambiguity Factor } \alpha = \frac{\textit{error (First local minimum)}}{\textit{error (Second local minimum)}} \tag{4.1}$$

As a special case, if the function has only one minimum, the error at the second minimum is defined to be infinity (yielding an ambiguity factor of zero).

Computation of the ambiguity factor in actual images is difficult. While it is easy to locate the first local minimum (the global minimum) in a vector, it is a bit harder to isolate the second minimum, because it is not necessarily the vector element with second-lowest value. Rather, it is the element with minimum value where the first derivative changes locally from negative to positive, outside the influence of the original minimum. Of course, in real signals there will be small perturbations in the evaluation function that should not be treated as actual minima, so these will have to be ignored in some way. We have created a useful (but not perfect) heuristic for finding the first n minima, that can be used to aid in the computation of the ambiguity factor. It works automatically by finding the minimum value, locally fitting a Gaussian curve to the values surrounding the peak, eliminating the influence of the Gaussian and iterating on the newly formed signal. The process repeats until the desired number of peaks or an evaluation function threshold is reached. This heuristic is discussed further in Section 4.4.1.

In summary, an ambiguous match occurs when a pixel's evaluation function exhibits multiple local minima. The presence of an ambiguous match does not necessarily mean the disparity estimate will be wrong, just that the potential for a false match exists.

4.1.2 Disparity Space

Having discussed the presense of an ambiguity at a single pixel, we now consider its effect on a scanline.

To understand how ambiguous matches arise we must consider the way stereo methods assign disparities. Recall that the goal of all filter-based stereo algorithms is to associate a unique depth value with each pixel in the input images. Most often a disparity value is assigned to a pixel; other values are possible (e.g., a pixel might be marked as occluded), but for now we only consider disparities. Theoretically speaking, each pixel in a stereo image can potentially be assigned any of a wide range of disparities. Stereo methods work by eliminating improbable disparities until only the most likely ones remain. To understand this process of elimination, we want to know the likelihood that each disparity will be assigned to each pixel. A useful visualization tool for this task is the disparity space.

The concept of *disparity space* is straightforward. Instead of associating a single disparity with each pixel, we associate a vector representing the likelihood of *all* disparities that might be assigned to it. This notion can be applied to a single pixel, a row of pixels, and even a matrix of pixels (i.e., an image). So while the disparity *map* for an image is 2D, the disparity *space* over a whole image will be 3D. To make it easier to visualize, we usually only consider the disparity space for a single pixel (by plotting the disparity evaluation function at that pixel, as in Figure 4.2), or for a single scanline (by creating a 2D image where intensity represents likelihood, as in Figure 4.3). Think of the disparity space as a visualization of the penultimate step in disparity computation; the disparity for a given pixel is determined by finding the disparity space vector element with minimum error.

Figure 4.3 contains an example of the disparity space matrix associated with a pair of image scanlines. The horizontal axis is indexed by Pixel Number and corresponds to the original scanline, and the vertical axis is the list of candidate disparities. Pixel intensities in Figure 4.3 correspond to the match error computed by the evaluation function; column 100 encodes the same information as in Figure 4.2. Dark regions in the image denote areas where the error is large, i.e., they represent those disparities that are least likely to become associated with the pixel numbered below. The disparity space makes the presence of ambiguities on a scanline easy to see. Any column with more than one bright spot represents a pixel likely to be involved in an ambiguous match.

The disparity space is a natural representation for search-based methods like dynamic

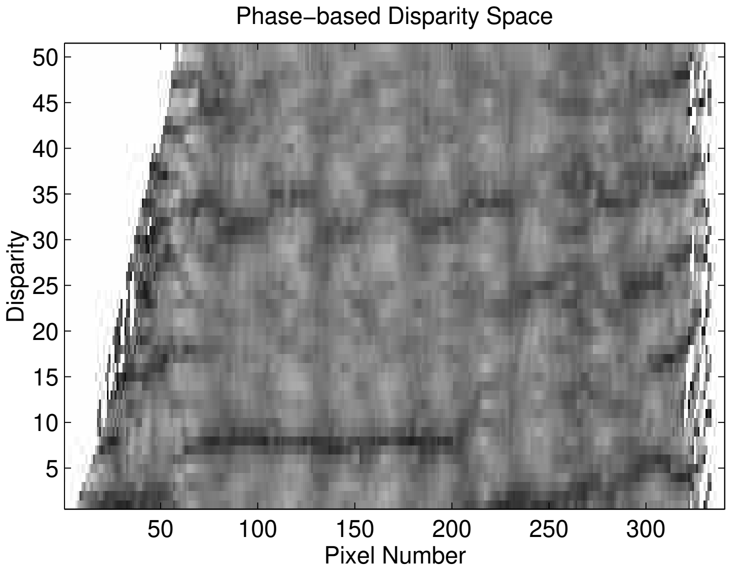


Figure 4.3: Disparity Space for row 170 of Figure 4.1 (with its pre-shifted counterpart), computed using our phase-based method with linear frequency samples. Darker pixels indicate lower matching errors, i.e., the most likely disparity estimates. Column 100 can be seen in an expanded view in Figure 4.2.

programming and our phase-based method from Chapter 3. In fact, the pseudocode for our algorithm presented in Table 3.2 computes exactly this information in its innermost loops. Note that by the time our phase-based method has constructed this image, all of the work in the frequency domain has been completed. The phase measurements have already been combined by the evaluation function to produce the likelihood that a particular disparity is appropriate. So the disparity space framework represents purely spatial phenomena, and can be applied directly to stereo methods that work exclusively in the spatial domain as well. We will elaborate on the benefits of the frequency domain in Section 4.3.3, but for now simply discuss the characteristics of the disparity space.

Once the disparity space is constructed, the usual way to generate a single disparity estimate for each original pixel is to find the minimum entry in each column, and call its associated disparity the best answer. We will often highlight these minimal disparities in disparity space plots by connecting them together across adjacent columns. In this way we can see not only the disparities for a particular scanline, but also the shape of the evaluation functions that led to their extraction. An estimate of precision can be provided as well, by measuring the curvature of the evaluation function around the minimum value (i.e., the second derivative of the evaluation function with respect to disparity, evaluated at that disparity at which the evaluation function has its minimum). Although one might consider using the raw evaluation function outputs as a measure of confidence and therefore precision, they are likely to be too variable to be of effective use. Simple changes in gain between the two cameras, or changes in object reflectance due to viewpoint change, could cause massive shifts in the amount of error, and therefore it is not in general robust enough to function as an independent precision estimator.

Disparity space images are not limited to search-based methods, they can also model the results of coarse to fine algorithms (e.g., the algorithm presented later in Table 4.1). The results are not quite identical to those of search-based methods, however. The main difference is that coarse to fine methods by their very nature do not compute an evaluation function at every possible disparity. Instead, they navigate through the disparity space in large ("coarse") steps, refining and evaluating only those subregions that meet some minimum criterion. As a result, the disparity evaluation functions associated with a given pixel are computed in completely different contexts, and therefore cannot be compared directly. So the argument above, that the disparity space was the penultimate step followed by minima extraction, is no longer appropriate. However, coarse to fine disparity spaces can still provide interesting visual information, especially when compared against a complete disparity space generated by a search-based method.

There are two complementary views of the disparity space for coarse to fine algorithms, each illustrated in Figure 4.4. The first view is identical to the one above, where we simply plot the value of the evaluation function between a pixel and a disparity. We cannot extract the minimum value to find the disparity in this representation, but it is still useful for visual inspection of the overall shape of the evaluation functions. The second view is the more appropriate one for coarse to fine methods. In this case the pixel intensities do not indicate the value of the evaluation function, but rather the number of the *highest scale* used to compute the evaluation function for a given pixel/disparity pair, with coarse scales numbered lower than fine scales. This plot has the same advantage for coarse to fine methods that the other plot had for search-based methods: the best disparities can be found by visual inspection, since the disparity returned by the algorithm will always be one of those searched at the highest scale. This plot also provides an explicit road map, demonstrating how the method chose to navigate through the detailed disparity space used by a search-based

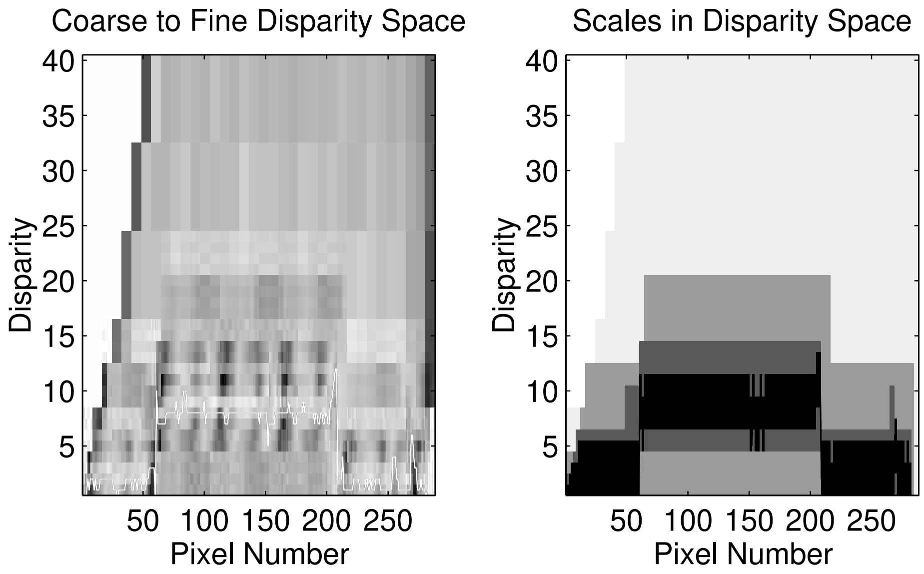


Figure 4.4: Coarse to Fine disparity space and associated scales; fine scales are darker than coarse scales. These results are also from row 170 of Figure 4.1, as are those in Figure 4.3.

method.

The actual content of the disparity space will depend on the evaluation function of each particular stereo method. It provides a way to visualize the search strategy used by that method, but should not be regarded as an inherent property of the image pair. Rather, it summarizes the interpretation of the image pair according to a particular method.

These disparity space representations provide the framework for our discussion of ambiguity.

4.2 Effect of Ambiguity on Stereo Methods

The usual effect of ambiguity on stereo methods is the production of outliers, sharp spikes in the disparity map. In the worst case, such as will be illustrated in Figure 4.7, *none* of the disparities is computed correctly.

4.2.1 JISCT Results

Out first illustration of the problem comes from a recent survey and evaluation of stereo methods, the JISCT Stereo Evaluation study (Bolles et al., 1993). Figure 4.5 is a stereo

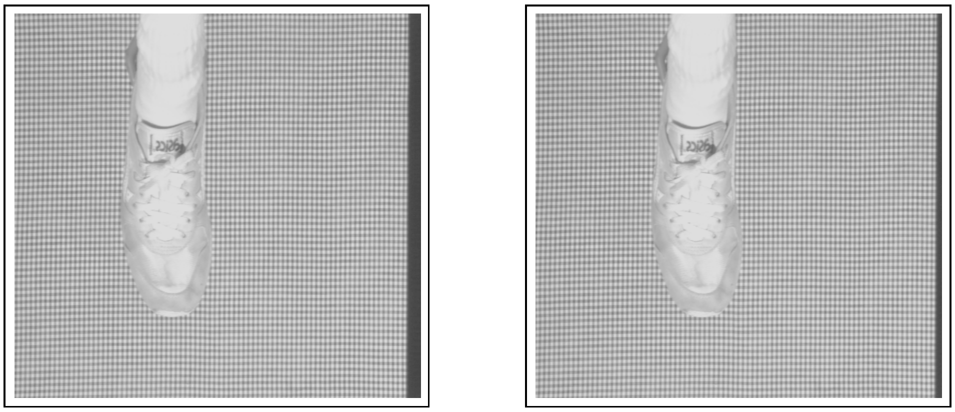
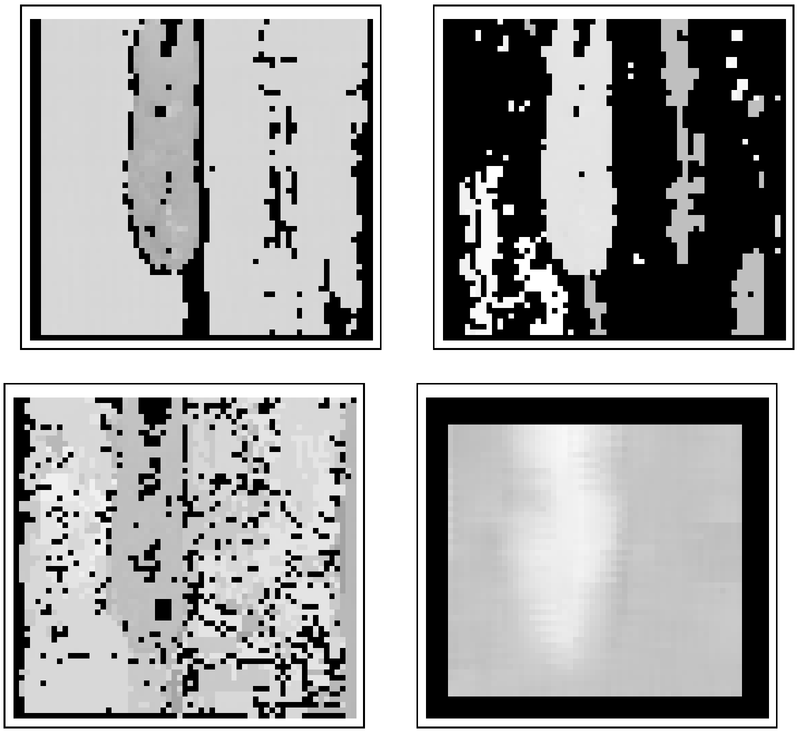


Figure 4.5: Shoe Stereo Pair. These are images SHOE-0 and SHOE-2 from the JISCT Stereo Evaluation study; each is 480x512 pixels[2].

INRIA 1

INRIA 2



SRI

Teleos

Figure 4.6: Results of several stereo methods on Figure 4.5 as presented in the JISCT Stereo Evaluation study. Clockwise from the top left are the INRIA 1, INRIA 2, Teleos, and SRI results; each is 59x63 pixels[2]. Black pixels were marked as unknown by the algorithms.

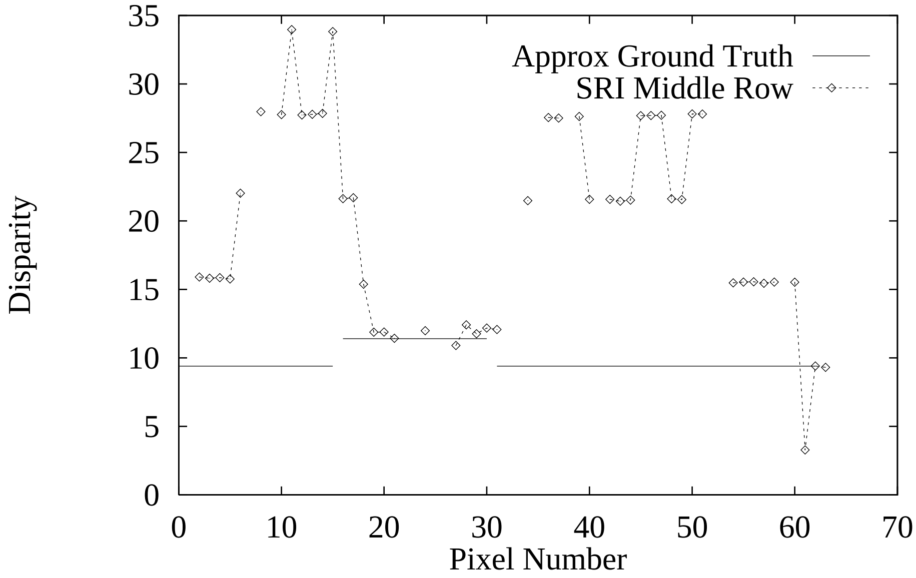


Figure 4.7: Detail view of SRI disparity results and approximate ground truth for the middle row of the shoe stereo pair disparities in Figure 4.6.

pair from that survey of a tennis shoe over a mostly-checkered background. Because the background is so regular, almost every pixel in it has multiple potential correspondents within a small area. Figure 4.6 shows the coarse disparity maps computed by several stereo methods in the study, illustrating the problems caused by the checkered pattern. In many of these results, the disparities computed within that pattern either jump around a lot (SRI and INRIA 2) or have significant gaps (all but Teleos).

These problems can be attributed to the repetitive nature of the input image. Each of the methods exhibiting spotty results uses correlation with 11x11 pixel filters as the final step (before interpolation) in assigning correspondence. Because this image is so regular, one match looks very much like another. Thus the background image (a carpet whose 3D shape is actually flat) appears to have many spikes in it. This is in spite of the fact that both methods explicitly try to eliminate such spikes: the INRIA methods employ morphological shrinking of the disparity map to remove outliers, and the SRI method uses multiple passes over the input data to eliminate unreliable matches. Yet you can see just how irregular the results are in Figure 4.7, which plots the SRI disparities for the middle row of the image against the approximate ground truth, generated by manual inspection. Not only are the results erratic, matching any of several candidate disparities, but most of the disparity estimates do not even match the correct piece of the pattern.

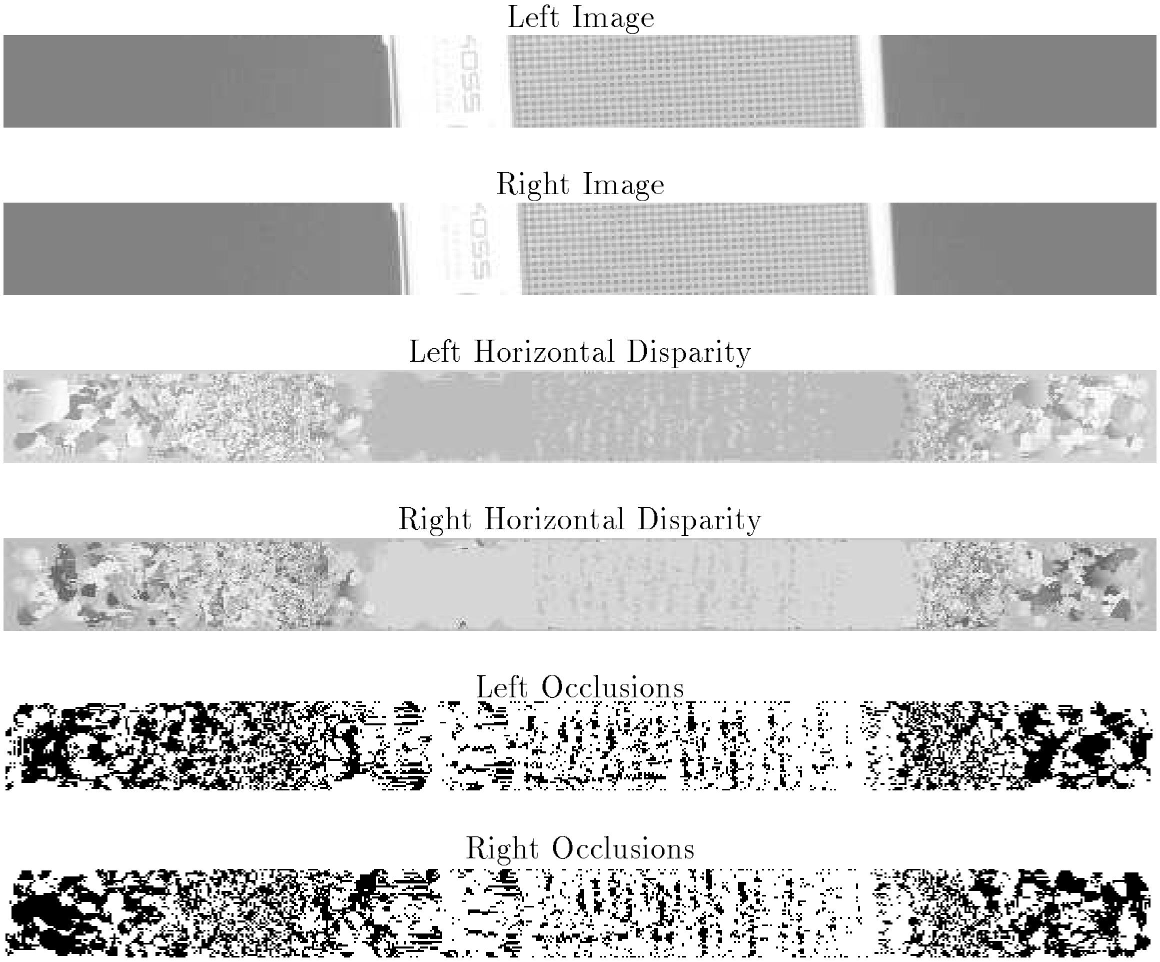


Figure 4.8: Results of Jones stereo method on middle region of speaker images.

Of the four methods in the JISCT study, only the Teleos algorithm was able to match the background successfully. Teleos' use of very large windows (from 25x25 to 90x90 pixels) helped them avoid getting trapped in the abundance of similar features at the finest scales, instead seeming to track stains in the carpet whose influence can only be seen at larger scales (Bolles et al., 1993). But this feature comes at a cost; since the search windows are so large, disparity is not computed at the pixels around the border of the image. Even more importantly, the shape of the tennis shoe (readily apparent in all the other methods) is greatly washed out, blurred by the extra-large search windows.

4.2.2 Jones' Method

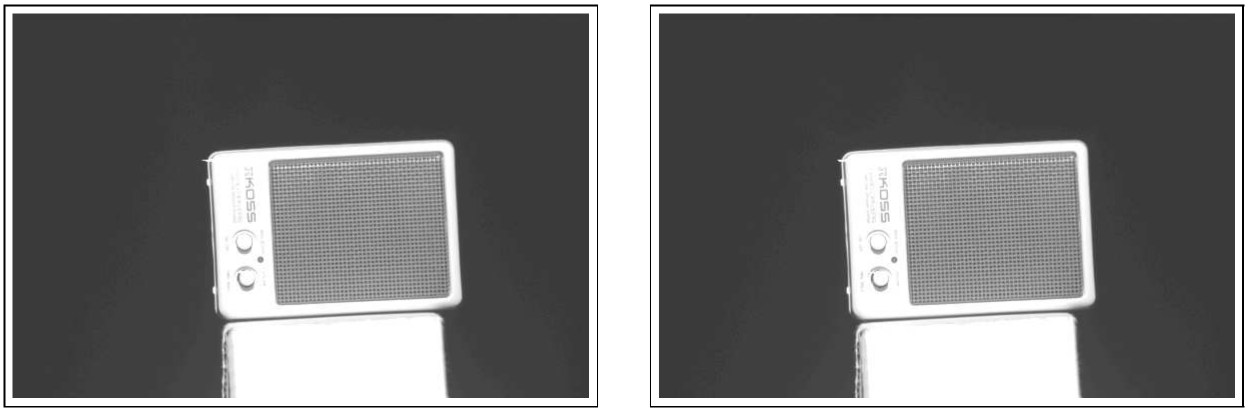


Figure 4.9: Speaker Image.

Another consequence of the disparity spikes caused by ambiguous matches is their effect on postprocessing techniques. For instance, Jones' stereo method (Jones, 1991) includes a special postprocessing step that addresses occlusions. The method compares disparities computed for the left and right images; if corresponding pixels do not have the same disparity, they are marked as occluded.[2] Given sharp disparity spikes, however, this step results in many spurious occluded pixels, as can be seen in Figure 4.8, especially in the area of the speaker grill. In this implementation of Jones' method, only the first pass and occlusion processing have been used.

4.3 Reducing the Ambiguity Factor

The goal of many stereo methods is to construct or modify the evaluation profile so that the Ambiguity Factor α is as small as possible. In the multibaseline method of (Kanade et al., 1995) this is accomplished by summing the evaluation functions across multiple image pairs. In a similar way, our phase based method sums evaluations functions from multiple *filter outputs*. However, the usual approach is to employ a coarse to fine refinement strategy. Coarse to fine methods smooth over the evaluation function, in an attempt to avoid falling into a local minimum.

In this section we will compare the coarse to fine approach with our phase-based filter combination method. We will see that although coarse to fine works well in some situations, in the presence of ambiguity its naive approach fails to resolve the ambiguity. We will

[2] A more robust method is outlined in Section 2.3.1 of this text.

Given: A pair of greyscale images L and R, search window size W, max disparity D, smallest resolution R, averaging filter *halve*.

Function *c2f* :

Matrix *prior, result* — *Disparity results from coarser level, and this level*

If (*resolution* / $2 > R$)
  prior = *c2f* (*halve* (L), *halve* (R), ...)
Else
  prior = $\lfloor \frac{D+2}{4} \rfloor \cdot$ I[*resolution*/2, *resolution*/2]

For each row r
  For each column c
    For disparity d = -$\frac{D}{2}$ to $\frac{D}{2}$
      $error(d) = \frac{1}{W^2} \sum\sum_{W}$ | left window (r, c) - right window $(r,\, 2\; prior[r/2, c/2] + d)$ |
    Pick d with minimum *error*, preferring values near 2 $prior[r/2, c/2]$ in a tie
    result (r, c) = d
return *result*

Table 4.1: Pseudocode for the Coarse to Fine algorithm.

demonstrate that our phase-based technique does a better job of compensating for this ambiguity.

4.3.1 Coarse to Fine Method

A common approach to overcoming the problem of ambiguous matches in two-camera stereo is to employ a coarse to fine (or image pyramid) search strategy. The intent is to limit the number of disparities checked at a given resolution; the fewer disparities checked, the lower the likelihood of an ambiguous match. An additional benefit is that a wide range of disparities may be checked for relatively small cost; the coarsest levels have the least data, yet span the widest disparity range in the original image.

To see how such methods compare with our phase-based approach, we have constructed a "typical" coarse to fine stereo algorithm. The method is typical in that it starts with a SAD correlation of the coarsest versions of the images, then iterates using only the results computed at the previous level to restrict the search at the current level. Pseudocode for this method can be found in Table 4.1. In practise we have applied this algorithm using a 3x3 Gaussian smoothing filter ($\sigma = 1$), search windows that are 5 pixels wide at each level, a 5 pixel maximum disparity at each level, and 20 pixels as the smallest resolution.

There is one aspect of this type of method in particular that distinguishes it from our phase-based method. Although both approaches consider the scale space decomposition of the original images, the coarse to fine approach marches through the scale space in a fixed order without regard to the content of the image. Decisions made at the coarse levels become prior assumptions at finer levels, and are taken at face value. Figure 3.10 demonstrates the scales used: the results from longer wavelengths are used to constrain those computed at higher wavelengths. In contrast, our phase-based method considers the entire scale space profile as a unit, selecting only those scales whose data are known to be valid, and combining them without regard to a predefined order. We will demonstrate shortly the benefit our phase-based method derives from this approach.

4.3.2 Coarse to Fine Results

The coarse to fine approach to image analysis is a popular one for two reasons. It is efficient, allowing large image regions to be searched at relatively low computational cost; and often it reduces errors, smoothing over small variations in the evaluation function.

An assumption implicit in the design of coarse to fine methods is that evaluation functions have an overall "bowl" shape, with a unique minimum. Minor variations (i.e., local minima) are tolerated as well, because they will be smoothed over by the processing at the coarsest levels. Another aspect of that assumption, not often stated explicitly, is that the evaluation functions at *all* scales must exhibit that structure; even the coarsest image is assumed to have a well-behaved evaluation function.

Figure 4.10 shows how the evaluation function evolves with each progressive level for a pixel in the wolf stereogram (Figure 4.1). Contrasting the final shape in Figure 4.10 with the evaluation function computed by the phase-based method in Figure 4.2, we see that the coarse to fine method was able to focus attention on the correct region in spite of the presence of the second and third local minima. So in this case the scale-based smoothing imposed by the coarse to fine method worked well, and allowed the correct local minimum to be extracted. You can view the final disparity map for this image (pair), as computed by the coarse to fine method, in Figure 4.24.

Problem with Coarse to Fine

Unfortunately, in some cases this model can lead to significant errors. The method worked well on the wolf stereogram because that image has a lot of texture at many scales; many details can be seen in the coarse images. But in some cases the coarse versions of images will *not* have sufficient structure. When that occurs, the method will still select a region in which to continue the search at a finer level, but that decision will be based on incomplete information. The coarse to fine strategy does not allow sufficiently for the possibility that a given level's evaluation function may not exhibit an obvious and unique minimum. Even though our implementation attempts to address this problem by preferring disparities near to the one predicted by the coarser level, it does not always succeed.

Consider the synthetic speaker grill in Figure 4.11. Although the original image contains much structure, and the correspondence between the two images seems obvious to a person, all of the interesting structure is washed out at coarser levels. Therefore, the coarse to fine search strategy cannot use the results from coarser levels to effectively constrain the search at finer levels. The resulting disparity image and coarse to fine disparity space for the middle row of the synthetic speaker grill illustrates the effect in Figure 4.12. Although the true disparity lies at a constant 23.52 pixels in the middle of the image, the coarse to fine method is unable to maintain the proper focus. Therefore most of its estimates are incorrect except at the very ends of the grill, where the grill and background colors cause a sharp edge to appear at all scales.

This effect is not limited to synthetic imagery, it occurs in real images as well, as can be seen in the results (and coarse images) from the Shoe stereo pair in Figure 4.13.

4.3.3 Solution using Phase-based Stereo

Our phase based method does a better job at addressing these problems. In this section we demonstrate that our method, as presented in Chapter 3, effectively reduces the ambiguity factor when the image pair contains enough information.

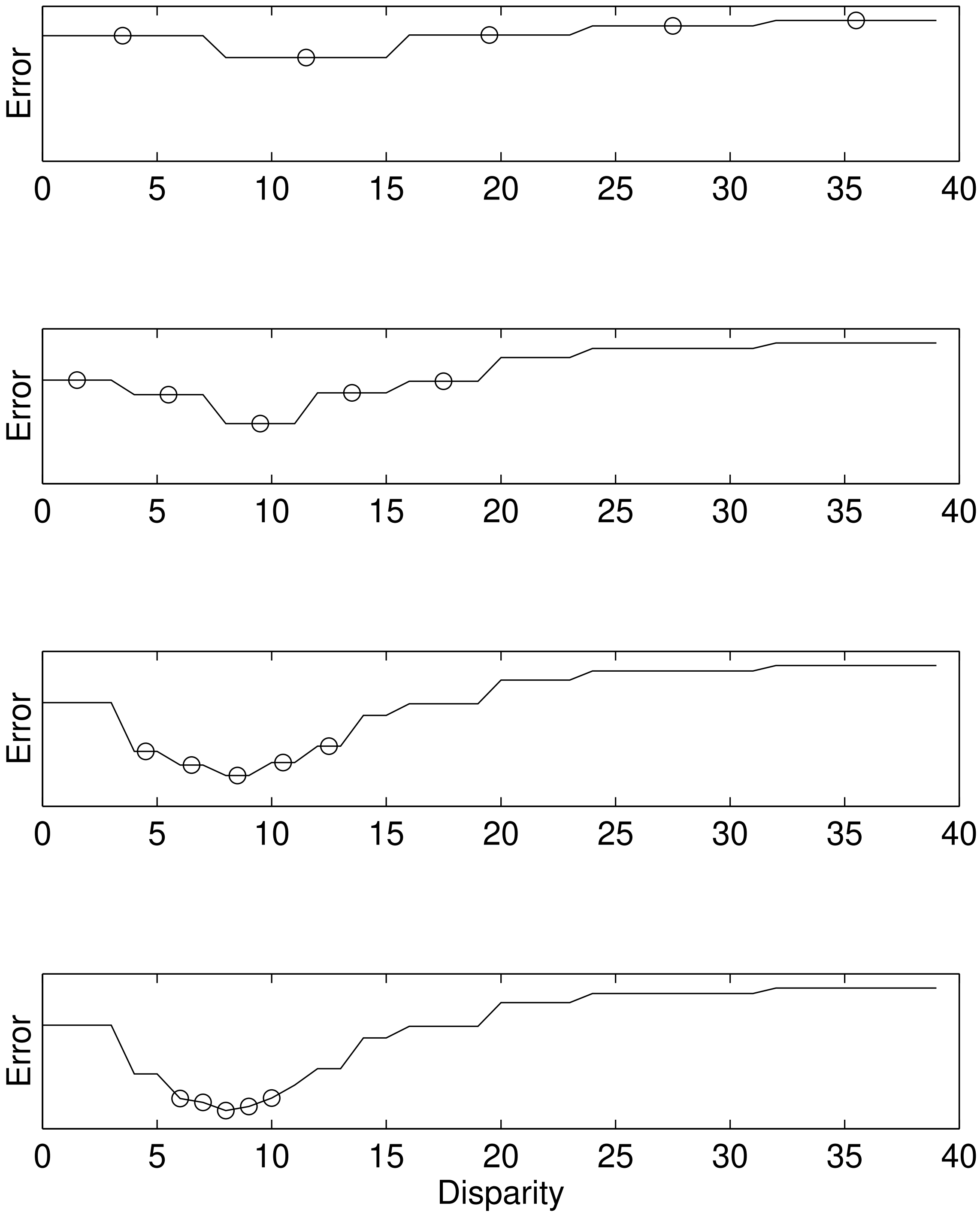


Figure 4.10: Evolution of the Coarse to Fine evaluation function at pixel (170,100) in Figure 4.1. The coarsest scale appears on top, the plots below demonstrate the successive refinement. Contrast the bottom plot with Figure 4.2.

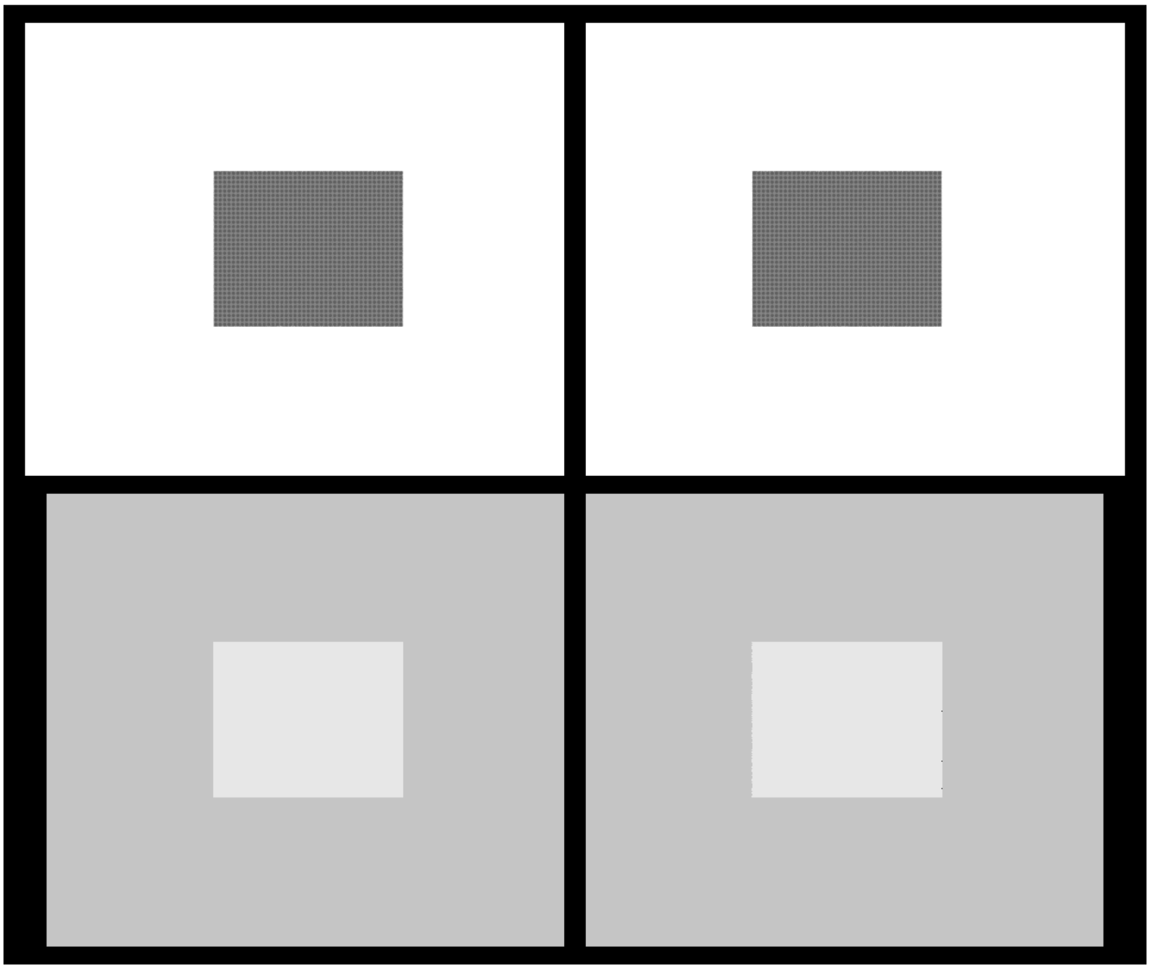


Figure 4.11: Synthetic Speaker Grill. Actual disparity is 18.3445 pixels, but the pattern repeats approximately every 3.8 pixels.

Our phase based method addresses the problem using the same tool as the coarse to fine method: a multiscale representation of the image. However, instead of enforcing an arbitrary ordering on the scales that are used, a more intelligent adaptive search is performed. In addition, results from different scales are combined according to the known relationship between the filters that created them. Because each scale generates a real-valued phase measurement, we have high-precision disparity estimates from *all* scales, not just the finest ones. These factors combine to yield a better behaved evaluation function in general, and in ambiguous areas in particular.

Confidence Estimates One of the advantages to the local spatial frequency framework is that each multiscale measurement includes a confidence estimate. In our phase-based

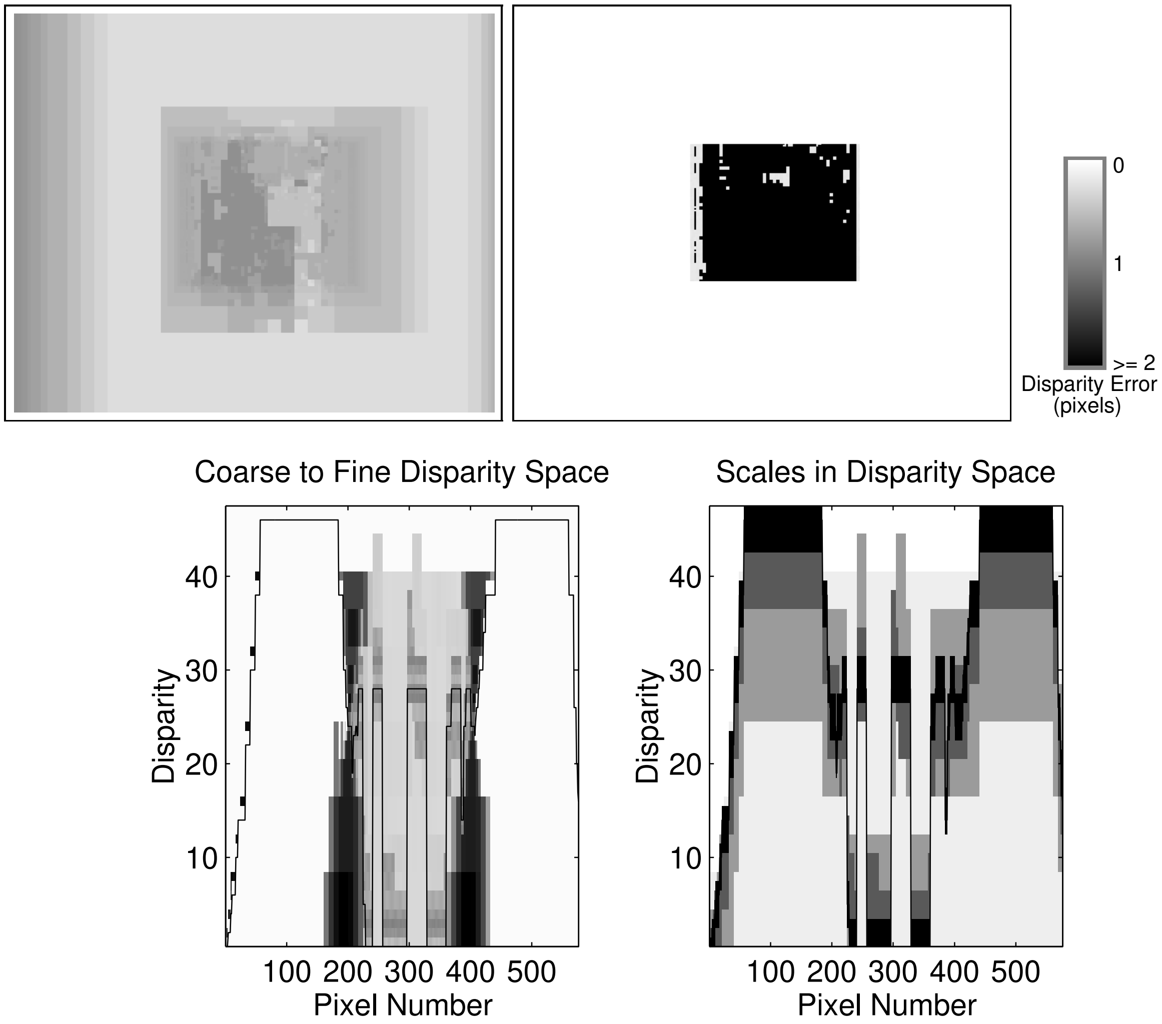


Figure 4.12: Coarse to fine results on Figure 4.11. Disparity map (upper left) has mean error 14.42 with $\sigma = 23.1308$ pixels, ground truth error image (upper right) maps all errors ≥ 2 to black. Lower plots are the coarse to fine disparity space (left) and scale space (right) for row 240 in the synthetic speaker grill. Ground truth for pixels 200–400 is 23.52 pixels.

method, the magnitude of the Gabor filter output is a measure of confidence in the phase value. More precisely, a combination of high magnitude and the constraint in Equation 3.10 provide a measure of confidence. This is an advantage over the coarse to fine approach in the spatial domain, which uses multiscale measurements without the benefit of an independent evaluation of their utility.

There are other phase based methods that use the coarse to fine structure, and therefore enjoy the benefit of a confidence estimate with their measurements. But they lack the ability

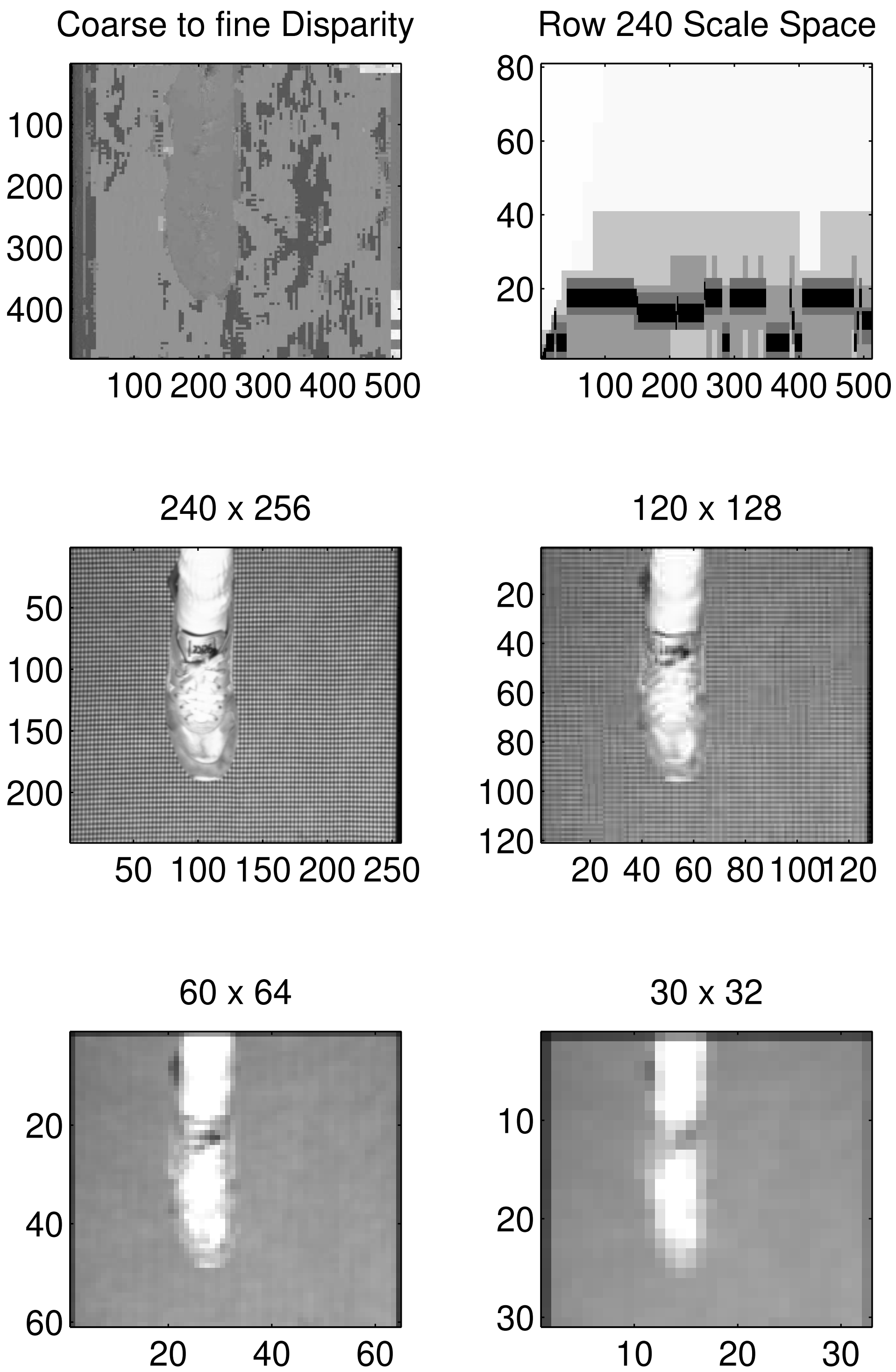


Figure 4.13: Coarse to fine results and images from the Shoe image pair.

to correct for data missing at the coarsest scales.

Scale Selection Thanks to the evaluation function, our method is able to use an *adaptive search through scale space.* To understand how this works, consider the error evaluation function in Equation 3.14. While the heart of the error computation is the difference between the ideal phase difference and that measured in the image, there is also a weighting term, which is the magnitude of the filter output. The effect of that term is to amplify the measurement errors of those filters with high magnitude. In other words, deviations around filters with low magnitude are less important than those with high magnitude, since they contribute less to the error.

Although this is an indirect argument, the point is that the filters with most reliable measurements will contribute the most to the matching error term. Thus the algorithm will choose the proper filters (i.e., scales) for the computation based on the characteristics of the images themselves, not because of an arbitrary requirement imposed by an inflexible search strategy.

Scale Independence Our method not only allows arbitrary scales to be selected, but also considers each scale independently from the others. Because each filter's contribution is merely summed with all the rest, there is no artificial order in which they must be evaluated. This is in contrast to coarse to fine methods, which rely on the coarsest scales to constrain potential mismatches at the finest scales.

High Precision Another advantage of the phase-based approach is the quality of the estimates produced at lower frequencies. Coarse to fine methods typically use lower frequency estimates only to restrict the range of disparity estimates checked at higher frequencies, but our method produces real phase measurements at each scale.

Examples

To understand why our phase-based method outperforms the coarse to fine approaches in instances of ambiguity, we will compare their evaluation functions. And we will use the scalogram to understand the reasons behind the improved shape of the phase-based evaluation functions.

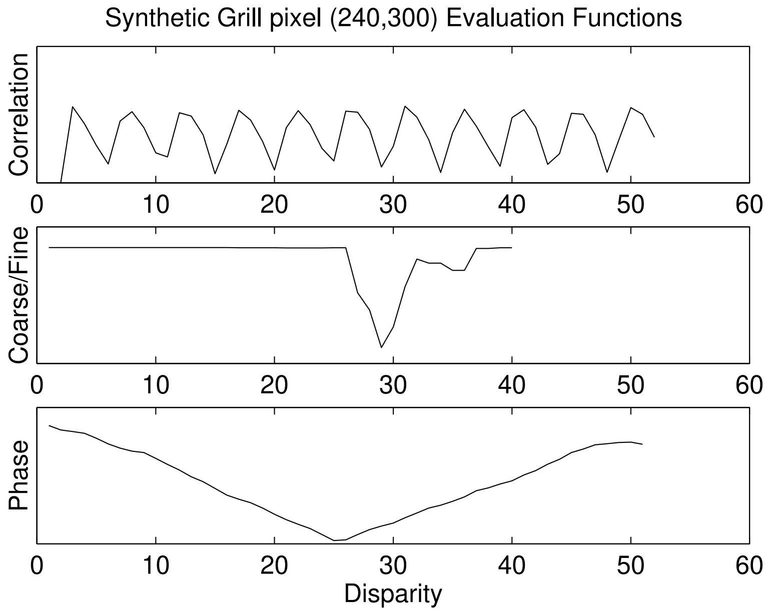


Figure 4.14: Synthetic Grill evaluation functions, illustrating the improvement of the phase-based method over raw correlation (too many minima) and coarse to fine (trapped in local minimum). Actual disparity is 23.52 pixels at pixel (240,300).

Synthetic Speaker Grill

The synthetic speaker grill image provides a nice test case. Its regular structure makes it a prime candidate for ambiguous matches in *any* stereo method, as can be seen in the evaluation functions for raw correlation, coarse to fine, and phase based stereo in Figure 4.14. In the figure, the raw correlation evaluation function profile exhibits ten local minima, all with approximately the same amount of error, resulting in an ambiguity factor of nearly 1. It seems clear that any local method such as this will be prone to ambiguous matches in the grill image pair, but what about a coarse to fine approach? Its evaluation function (also in Figure 4.14) has a better profile, with a unique global minimum and low ambiguity factor, but still has problems. The evaluation function has more than one local minimum, but even worse, the unique global minimum found by the algorithm is *not* the correct disparity. Both of these algorithms were fooled into making a false match by the numerous candidate matches, a problem which their search strategy was unable to model successfully.

Our phase based approach not only finds the correct solution, but also exhibits an evaluation function profile that would be considered ideal by any stereo method: a virtually unimodal function with a single minimum located at the correct disparity. And as the Dis-

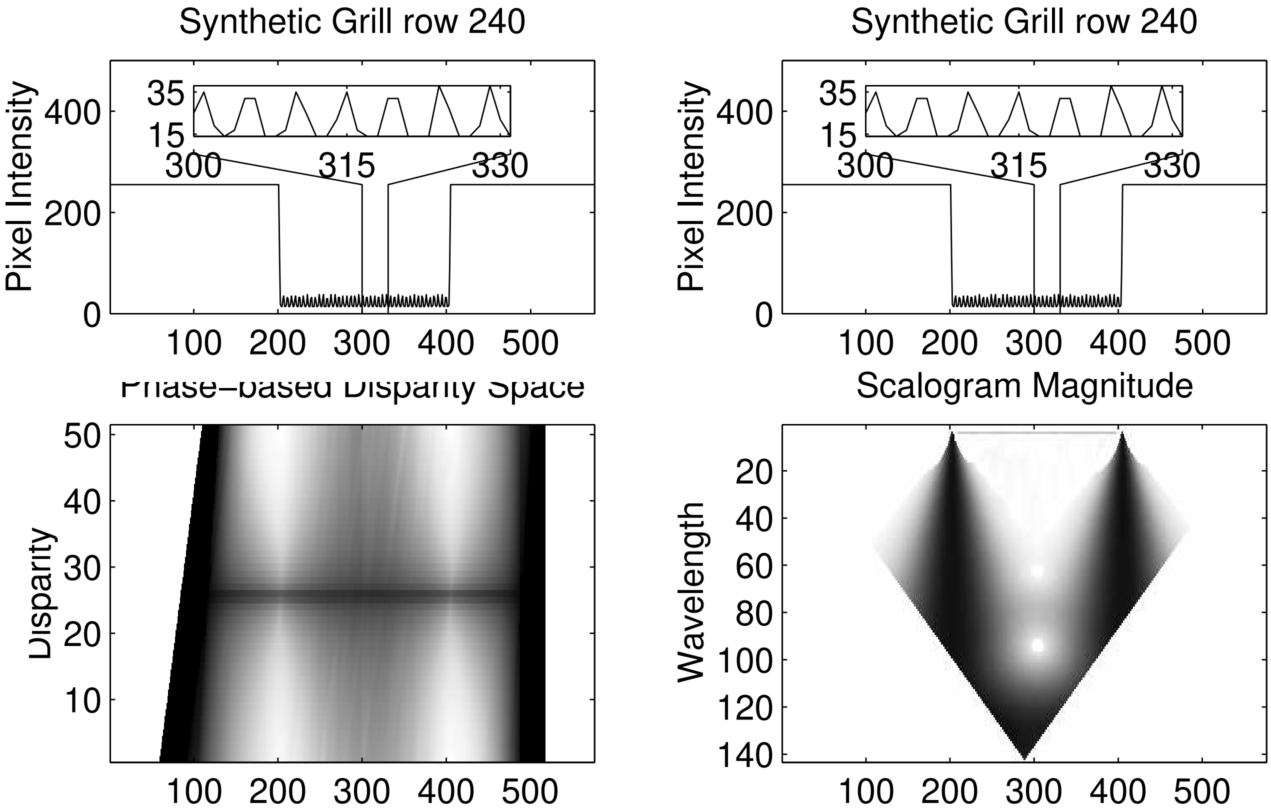


Figure 4.15: Synthetic Grill Phase-based Disparity Space and constraint-filtered Scalogram for line 240. The raw pixel intensities appear above (the same scanline is plotted twice), and illustrate the effect of the grill edges on the plots below.

parity Space in Figure 4.15 shows, *every* pixel in this row has a similarly-shaped evaluation function pointing to the correct disparity. How was our method able to succeed where the others failed?

The explanation can be found in an analysis of the left magnitude scalogram for this row, shown in Figure 4.15. By far the most prominent features are the dark vertical columns that correspond to the edges of the speaker grill. Their presence is to be expected; the Fourier transform of a step edge (e.g., the light/dark transitions in the original image) has energy content at all frequencies. Another expected feature is the horizontal bar at the top of the scalogram. This bar extends across the width of the speaker grill, and its row number corresponds to the period of repetition of the holes in the grill (about 3.8 pixels). The presence of a specific peak in each column like this is exactly what you would expect from an image intensity profile that looks like a sinusoid. But the key to understanding why the evaluation function maintains such a useful profile over the whole row is the dispersion of the dark columns that occurs at the edges of the grill.

Recall that our phase-based stereo method compares columns in the left and right scalograms. This dispersion results in higher magnitudes, and therefore more reliable phase measurements, not only at the edges of the grill but also at low frequencies throughout the central grill region. If only the highest frequencies in that region were considered, then the evaluation functions for the center pixels would look much like the raw correlation profile in Figure 4.14, thanks to the phase wraparound problem mentioned in Section 3.6.1 and illustrated in Section 2.2.1. But in this case the analysis filters with lowest frequencies were wide enough to include the strong influence of the step edge at the borders of the region. So when the columns that represent pixels in the middle of the grill are considered, the analysis also includes nonlocal information about the full extent of the similarly-textured region; the filters in the central columns know about the edges of the grill too. Therefore, our approach to multiscale analysis is able to combine the very high and very low frequency measurements effectively, eliminating the ambiguity faced by other methods.

This analysis also clarifies the reason the coarse to fine method failed to resolve the ambiguity. Even though there is good information at coarser scales (i.e., low frequencies), there is *no* information at medium scales, as the white region in the middle of the scalogram in Figure 4.15 demonstrates. And since the coarse to fine method requires that some decision be made at *each* scale, it is not surprising that decisions made in the absence of useful information at middle scales would lead to incorrect results.

Shoe

Our method is not foolproof, however. Consider the results of the phase-based method on the Shoe image pair. The results look quite encouraging; the shoe disparities look good, the background color is solid. But the background pattern is a repetitive checkerboard, and the disparity found by our method is *not* the actual disparity; this is shown graphically in Figure 4.16.

In fact this image pair was designed to produce ambiguity, as Figure 4.19 illustrates. The fact that real imagery can exhibit inherent ambiguity like this forces us to redefine our notion of disparity at a pixel.

4.4 Modeling Ambiguity

Sometimes it is simply not possible to reduce the degree of ambiguity present in an image. In such situations it is best to model the ambiguity, rather than attempt to correct the evaluation function by assuming no ambiguity exists. In this section we present a framework for such an analysis, and demonstrate its potential for improving disparity estimates.

4.4.1 Extended Disparity Representation

Our ambiguity model requires us to extend the traditional notion of depth values recovered by a stereo method. The usual model for depth at a pixel is a single disparity estimate, possibly with an indication of the precision, or variance, of the result. Unfortunately, this model does not include any information regarding the accuracy of the estimate. In fact, it presumes an "all or nothing" approach to depth measurement; either the measurement exists and is known to a certain precision, or no depth estimate is known (e.g., the pixel is occluded). Thus the model implicitly assumes that all evaluation functions will have a single unique minimum. As this chapter has demonstrated, e.g., in Figure 4.14, there are times when this simplified model is not sufficient.

The problem is that this model assumes a unique disparity can always be found. While the 3D point imaged at a given pixel will certainly have a unique depth, there may be many plausible disparities under a particular 2D evaluation function. The only way to incorporate

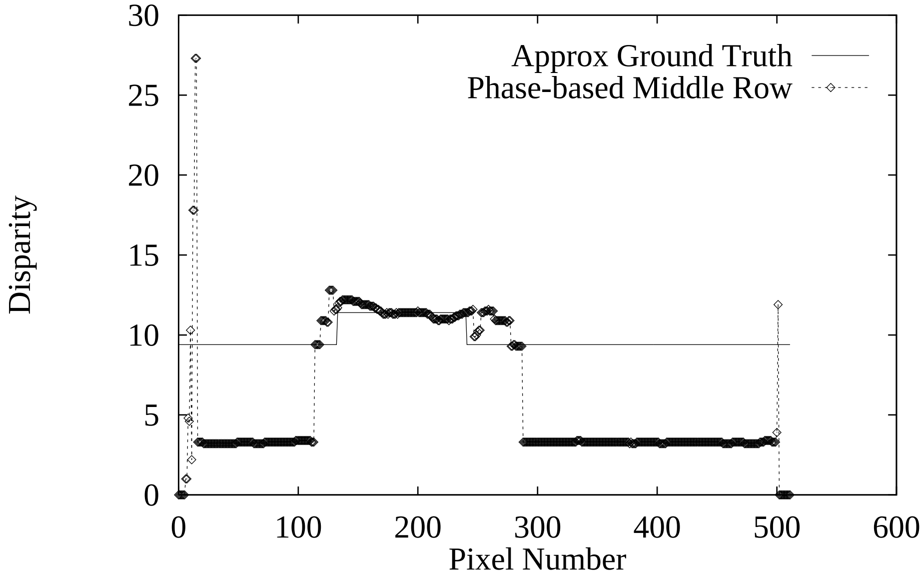


Figure 4.16: Detail view of phase-based disparity results and approximate ground truth for a middle row of the shoe stereo pair. Background disparities are consistent, but incorrect. Figure 4.21 contains the complete disparity space for this scanline.

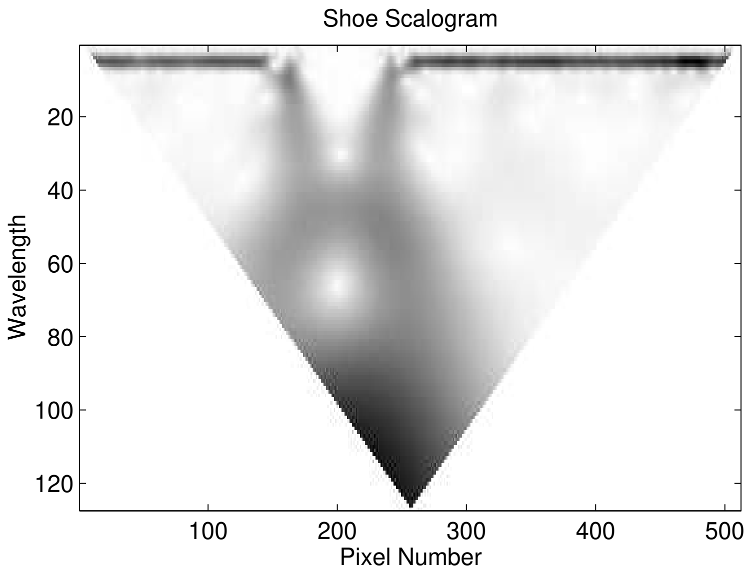


Figure 4.17: Image Scalogram for the middle row of the left shoe image.

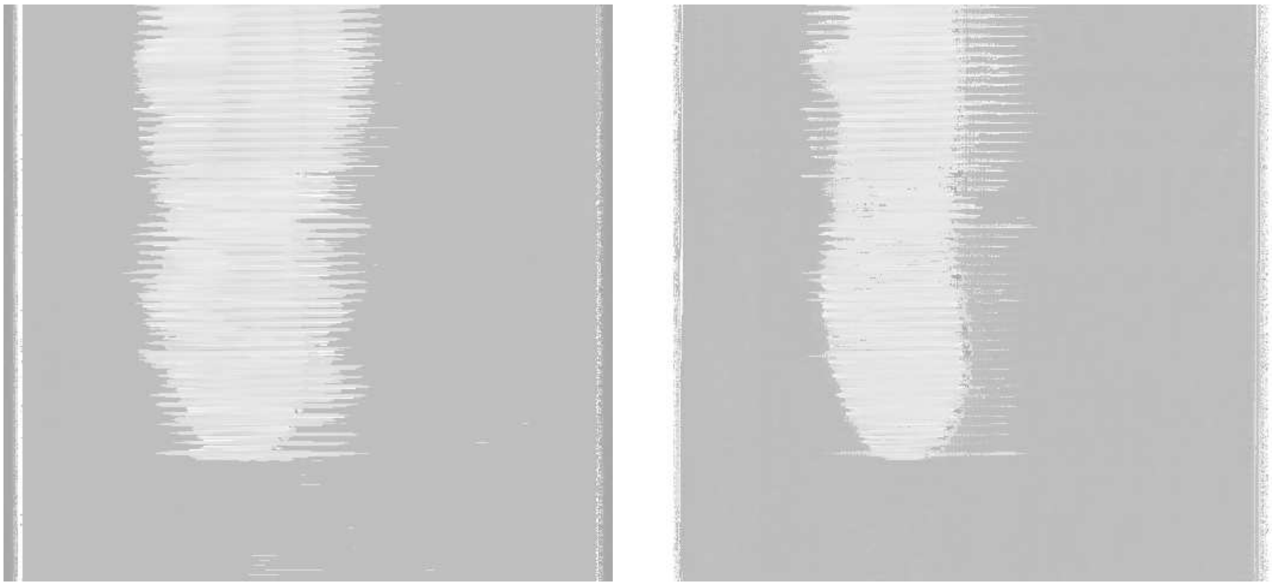


Figure 4.18: Phase-based disparity maps for the shoe image pair. Left map is the result from using no constraints, right map is the result from using the heuristic in Section 3.4.3.

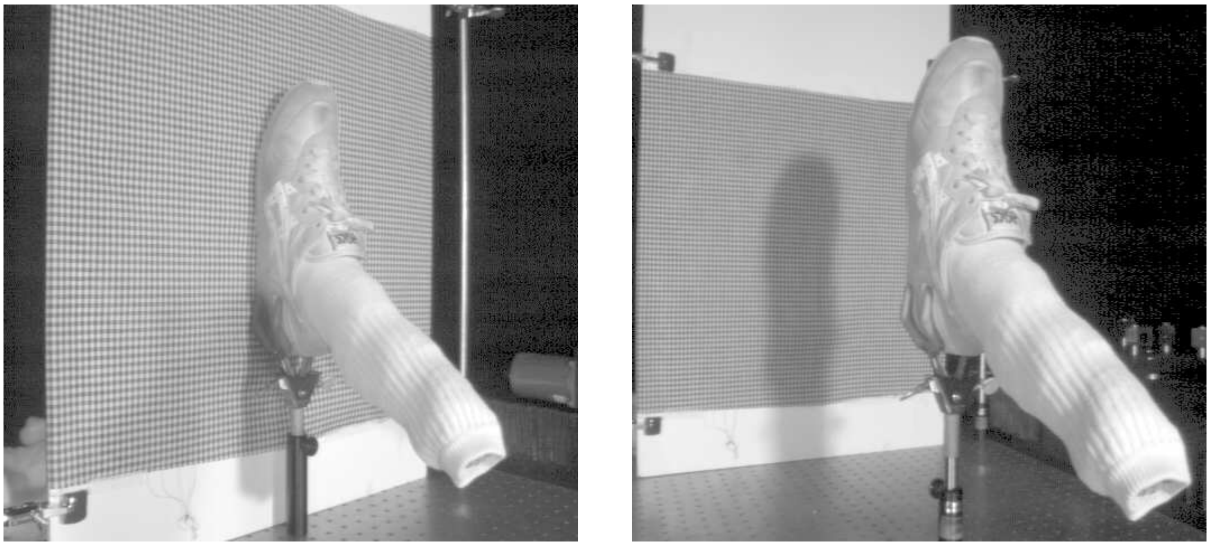


Figure 4.19: A view of the deviousness used in the construction of the Shoe image pair. Although the images were taken of the shoe flat against the texture (left image), the fact that the actual disparity is larger than the period of the checkerboard pattern causes a purely local search to reconstruct the scene with the shoe "floating" above the textured background (right image). Reproduced with permission from Kanade.

multiple candidate disparities under the unique disparity model would be to decrease the precision of the estimate by many pixels so that all candidates are included. That would be a needless waste of information since each estimate is likely to be known rather precisely, often to subpixel precision. A better model would allow a pixel to have *several* disparity estimates, each with its own precision and relative likelihood. More formally,

$$Estimates \;=\; (\;\langle disparity,\; variance,\; error\,\rangle\; \ldots\,) \tag{4.2}$$

where *disparity* is the best estimate at a local minimum, *variance* is the curvature of that minimum, and *error* is the evaluation function value at that minimum, expressed relative to the other estimates at this pixel using the formula:

$$Error \;=\; \frac{error_n}{\max_{i \in \mathrm{PeakIndices}} error_i} \tag{4.3}$$

The *PeakIndices* are those disparities whose evaluation functions exhibit local minima. Their automatic extraction is in general very difficult, but we have constructed a heuristic approach that works well enough to demonstrate the principle. By applying the following heuristic to the evaluation function of single pixel, e.g., Figure 4.2, not only the global minimum but *all* useful minima may be extracted. In this way multiple disparity estimates may be generated at a pixel.

Automatic Estimate Extraction

The locations of local minima in a sampled function can be identified using the heuristic peak finder first described in Section 3.4.3. Recall that the method works by locating the global maximum, extracting a window of values around the maximum (up to the point where the sign of the second derivative changes), fitting and subtracting a Gaussian to those points, then iterating on the newly subtracted signal. The intuition behind it comes from the observation that the same process that creates the local maximum will have residual effects nearby, effects that diminish with distance from the peak. This occurs in the frequency domain application of Section 3.4.3 because the frequency response of the Gabor filters has a Gaussian envelope, so nearby overlapping frequencies will indeed see diminishing effects that fall off as a Gaussian. In this new domain of evaluation functions, the assumption is that the evaluation function error will increase monotonically when moving away from the minimum.

A few adjustments do need to be made to accommodate the new domain. Since the original method finds local maxima, and we are now interested in minima, the evaluation function values must first be negated and translated to have a nonnegative minimum value before calling the peak-finder. Also, while the magnitude profile of overlapping filters tends to vary smoothly, in this case the evaluation function profile may have more abrupt transitions, and generally will be subject to more noise. Therefore another threshold is introduced, to enable the method to ignore small perturbations in its input. Whereas the previous method expanded the window around the maximum until the second derivative became greater than zero (i.e., changed sign), for this application we will set the threshold slightly above zero, e.g. 0.5. The effect of this threshold is to allow the support window to grow larger in the presence of noisy data, which should enable a better Gaussian fit. Finally, in order to avoid spurious results from the smaller peaks that result from the subtraction step, we mark all values that contribute to the Gaussian fit as unusable peaks. Should one of those pixels be found to be the maximum, the iteration of the peak-finding procedure will terminate.

Figure 4.20 illustrates this heuristic procedure on two examples. Regions of the function that contributed to the Gaussian fit are marked with dashes in the figure, and each extracted maximum is highlighted with a dark circle.

Output from this heuristic provides exactly the multiple estimates demanded by Equation 4.2. The location of the peak indicates the *disparity*, the curvature of the evaluation function provides the *variance*, and given a complete set of peaks the *error* can be computed using Equation 4.3. In the next section we apply this heuristic to an image pair to illustrate the benefits of the representation.

4.4.2 Demonstrating Improvement on an inherently ambiguous image

Given the locations of the local minima in a pixel's evaluation function, multiple precise disparity estimates may be associated with that pixel. Although multiple disparities can model the results of stereo methods more accurately, in practise it is extremely difficult to put them to good use. The problem lies in the generalization from pixels to surfaces: if all possible combinations of even as few as two estimates were considered, the number of possible surfaces in a single *scanline* would be 2^{512}, i.e., beyond astronomical. We will not propose a general solution to this problem, but rather will demonstrate that useful results

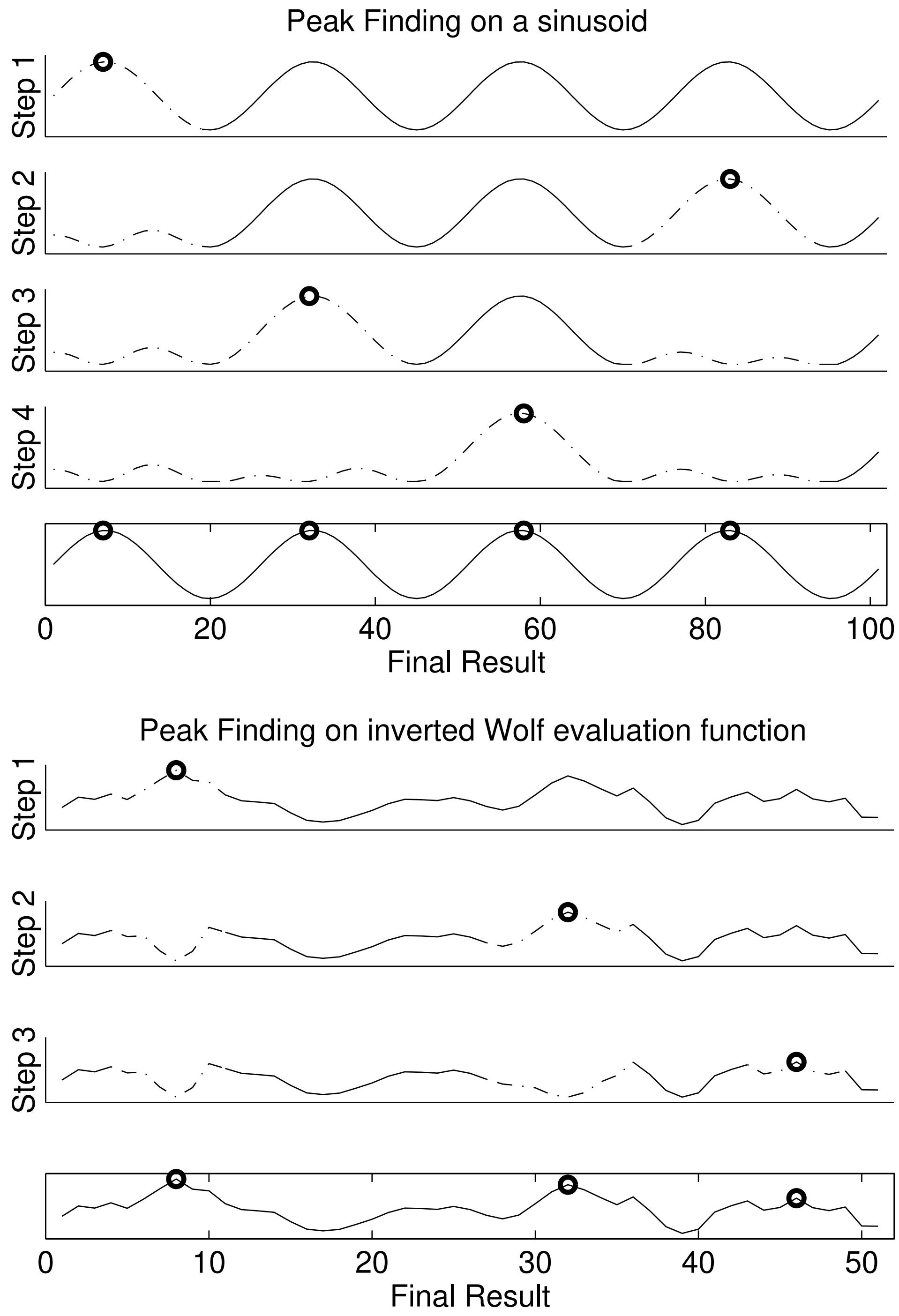


Figure 4.20: Illustration of the peak-finding heuristic from Sections 4.4.1 and 3.4.3 on a sine wave (upper) and the inverted Wolf Evaluation Function from Figure 4.2 (lower). Dashed lines indicate the region in which a Gaussian is fit and then subtracted.

may still be obtained.

We will use the Shoe image pair for this demonstration, since even our phase-based method was fooled by its inherent ambiguity. Figure 4.21 illustrates the phase based disparity space for the central scanline of the image pair, with the scanline plotted above for comparison. In addition, local minima extracted by the heuristic (using a peak threshold of 70% of the maximum) have been identified and plotted directly over the disparity space image in Figure 4.22. Only the peak locations have been plotted, their variance and error components are implicit in the disparity space image. These local minima make any inherent ambiguities plain: when a column exhibits more than one local minimum, the corresponding pixels will have a non-zero ambiguity factor.

The nature of the overall ambiguity becomes clear by inspecting the connectivity between adjacent columns. All columns in the area representing the shoe (pixels 150–250) have a single minimum at the correct disparity, and are thus clearly unambiguous. The area of the checkerboard pattern has lots of peaks, though; individual columns have as few as two or as many as nine potential matches within the given disparity range (0-50 pixels). The majority of those peaks fall into disparities about 6 pixels apart: this corresponds nicely to the period of the checkerboard in the original image, which agrees with our intuition that the self-similar checkerboard texture makes pixels in the background prime candidates for ambiguous matches.

The utility of this representation becomes clear when you contrast this combined minima plot (Figure 4.22) with the disparity results of the phase based method (Figure 4.16) and of SRI's method (Figures 4.7 and 4.23). From manual inspection of the original image pair in Figure 4.5 (especially the black stripe on the right hand side that indicates the end of the texture), we know that most of the actual image disparities lie around 10 ± 2 pixels. The incorrect results from both methods are easily explained as alternative local minima in the disparity space that were chosen because the evaluation function rated them more favorably. By plotting all of the local minima, we are more likely to find the actual disparity along with other potential match points.

Once all the candidate matches have been found, we can accurately describe the match quality over the entire scanline. Columns with a single match can be interpreted as before, where a single disparity value represents the most accurate reconstruction possible. In areas with multiple candidates, techniques other than finding the raw evaluation function

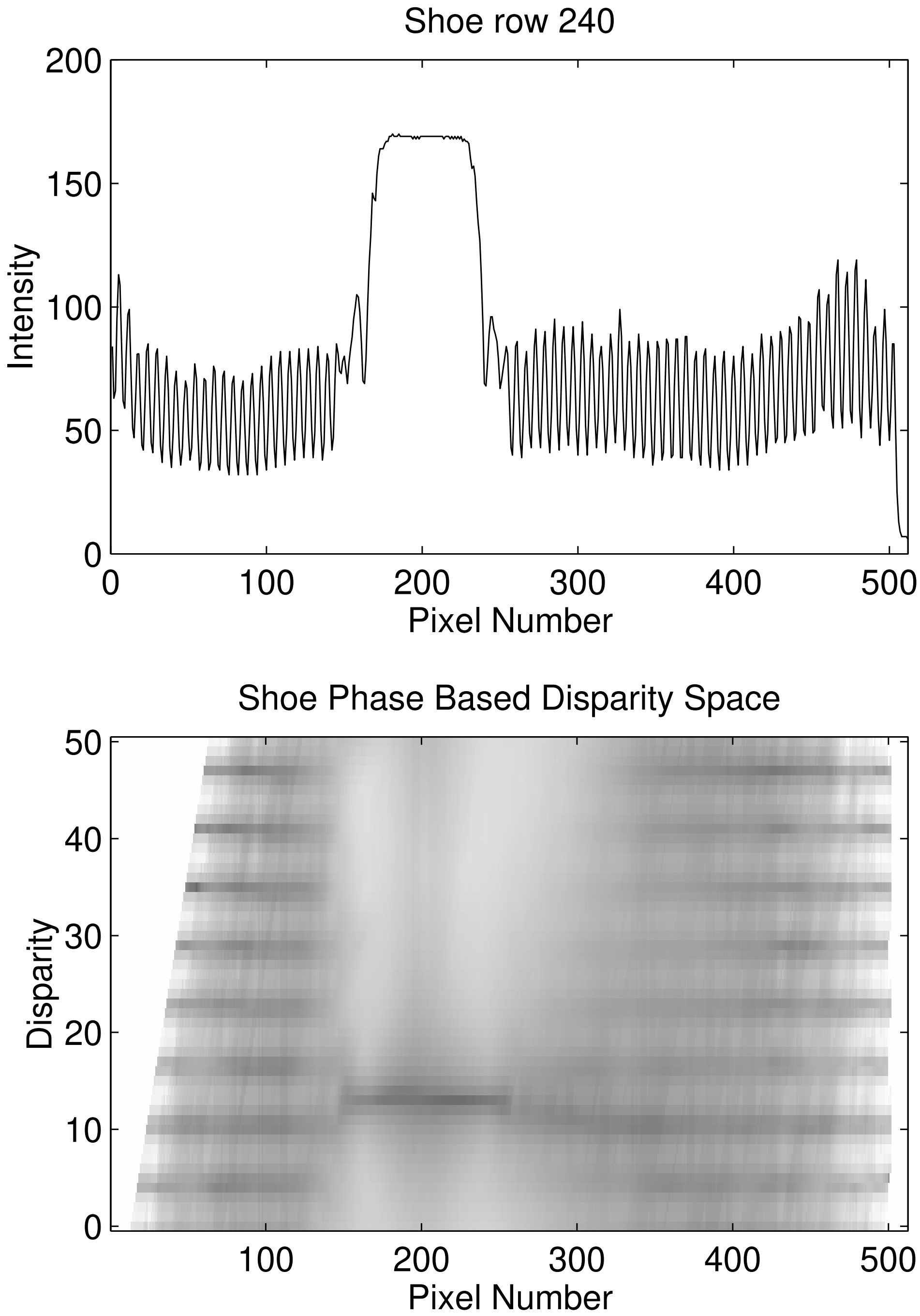


Figure 4.21: Shoe Disparity Space. The upper plot is row 240 from the Shoe image pair, the lower plot is the disparity space computed by our phase-based method.

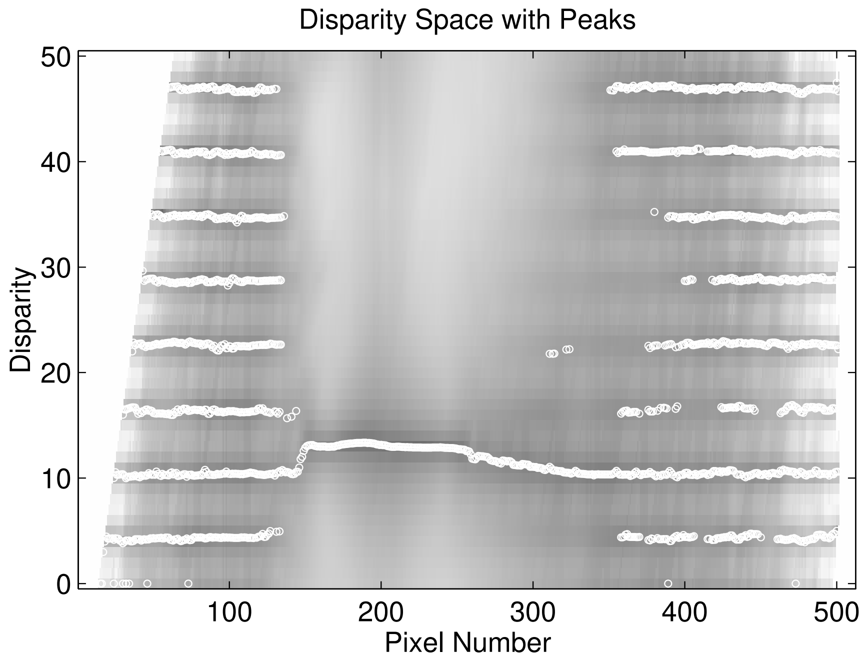


Figure 4.22: Phase-based Disparity Space with Peaks. Peaks computed using the heuristic have been superimposed on the disparity space from Figure 4.21.

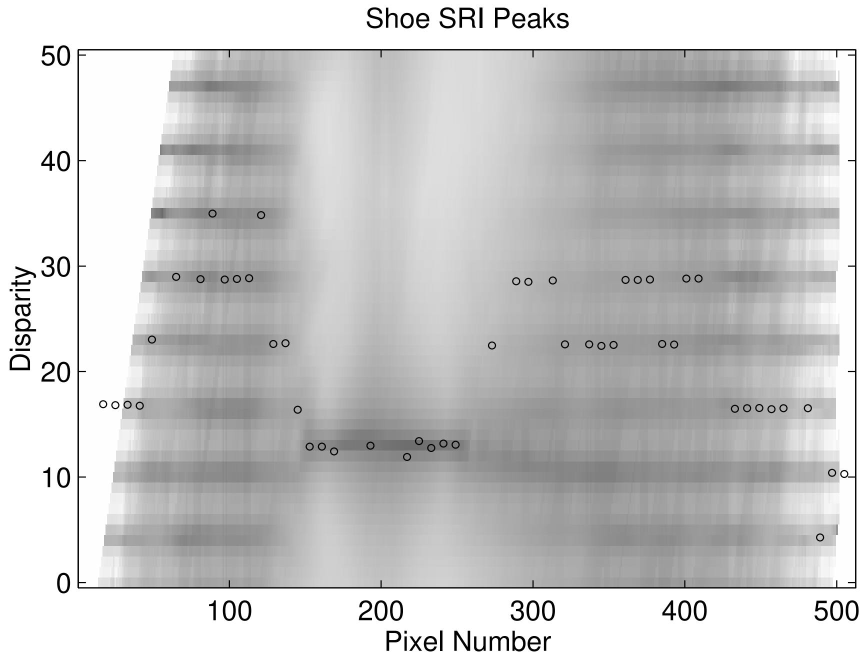


Figure 4.23: SRI Disparity over Phase-based Disparity Space. The seemingly random results from Figure 4.7 actually fit nicely into the local minima of the phase-based disparity space (the SRI disparity space was not available).

minimum can help eliminate false targets. For example, morphological operations on a binary "peaks-only" image could be used to group similar disparity estimates together and eliminate outliers. In the shoe example, such a procedure might find that the background region has eight possible interpretations that can be grouped into three categories: a plane that lies much closer to the camera than the shoe, nearly the same distance as the shoe, or much further than the shoe. Such a procedure might average the *error* terms of the candidate matches to present those multiple interpretations in a particular order. By grouping the multiple candidate matches together at adjacent pixels, we reduce the complexity of the model from an astronomical number of independent pixel disparities down to a more manageable eight potential surfaces.

A similar approach has already been demonstrated in (Zitnick & Webb, 1995). The core of their stereo algorithm groups match candidates into surfaces, then chooses the surface with the greatest number of points as the correct one. However, their method currently requires that a pattern be projected into the scene, so it is not a passive stereo method. Further details of their method are beyond the scope of this work, but the interesting point is that nearly all prior work on stereo has focused on improving the shape of the evaluation function, while Zitnick and Webb eliminate the evaluation function entirely, relying instead on pixel to pixel matches (i.e., a 1x1 pixel evaluation window). This chapter has presented a compromise; improve the evaluation function, *and* work with the extracted potential mismatches.

4.5 Summary

The problem of ambiguous matches, or false targets, can greatly reduce the accuracy of a stereo vision system. We have shown that the usual approach to alleviating the problem, a coarse to fine refinement strategy, imposes some (perhaps overly) strong requirements on the stereo images. Our phase-based method relaxes those requirements, and is therefore able to handle a wider variety of otherwise ambiguous images. We have also proposed a generalized disparity model that explicitly represents multiple candidates. This model allows higher level functions to reason more accurately about the structure of the image.

Figure 4.24: The structure embedded within Figure 4.1. This image was computed by applying our coarse-to-fine stereo method with 5 pixel window and 5 pixel max disparity per level to the original 340x340 image and a copy of the original shifted by 58 pixels.