Modeling Foreshortening in Stereo Vision using Local Spatial Frequency

Mark W. Maimone Steven A. Shafer
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213-3891

Abstract

Many aspects of the real world continue to plague stereo matching systems. One of these is perspective foreshortening, an effect that occurs when a surface is viewed at a sharp angle. Because each stereo camera has a slightly different view, the image of the surface is more compressed and occupies a smaller area in one view. These effects cause problems because most stereo methods compare similarly-sized regions (using the same-sized windows in both images), tacitly assuming that objects occupy the same extents in both images. Clearly this condition is violated by perspective foreshortening.

We show how to overcome this problem using a Local Spatial Frequency representation. A simple geometric analysis leads to an elegant solution in the frequency domain which, when applied to a Gabor filter-based stereo system, increases the system's maximum matchable surface angle from 30 degrees to over 75 degrees.

1 Introduction

Object surfaces are rarely viewed head-on in both images of a stereo pair. Instead, they may appear more compressed in one image due to perspective foreshortening, as in Figure 1. When a surface has a textured appearance, this effect makes matching the two images very difficult, since its appearance differs so much between the two images. This leads to poor results from area-based stereo matching techniques.

In this paper we develop a model of perspective foreshortening that enables us to quantitatively predict its effect on stereo image pairs. We present two equivalent forms of a correction factor that allow us to reason about foreshortening effects in both 3D world coordinates and 2D image coordinates. We show how to improve the accuracy of phase-based stereo matching systems using this information, and demonstrate its application to a particular Gabor filter-based stereo system. Applying the correction factor to this system increased its maximum matchable surface angle from 30 degrees to over 75 degrees.

2 Related Work

Several phase-based stereo methods have been described in the literature [3] [12], and a review of the more popular variations can be found in [4]. Although some of these mention foreshortening as an issue, none has explicitly modeled or corrected for it.

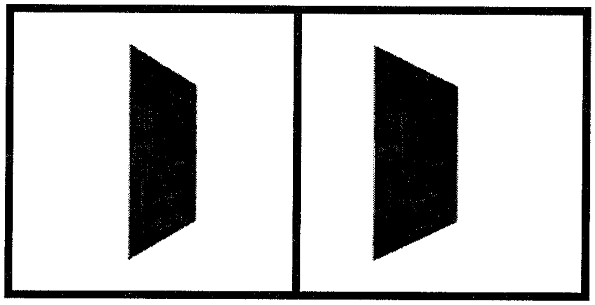


Figure 1: Stereo pair illustrating the effects of foreshortening; image compression, differing extents.

There have been a few attempts at modeling foreshortening in the context of stereo matching. Jones and Malik [5] attempted to apply local spatial frequency to the problem, but assumed orthographic projection and affine warping. Belhumeur [2] addressed the problem in the spatial domain, but his method requires an estimate of the disparity derivative, an inherently noisy estimator. Panton [10] also dealt with foreshortening in the spatial domain but assumed that the average global depth was fixed and the surface slant was bounded. The variable window method of Kanade and Okutomi [6] implicitly addresses foreshortening in the spatial domain by allowing corresponding windows to have different sizes, but is intended to function as a high-precision refinement technique: without proper guidance from other sources it tends to get stuck in local minima and flatten out sloped surfaces.

Local spatial frequency has already been identified as a valuable tool for modeling surface shape and segmenting multiple textures in a single image [7] [9]. These approaches use filter magnitude in the frequency domain as the feature of interest, and require either that the surface textures exhibit specific properties (e.g., periodicity), or that they be viewed directly head-on.

Local spatial frequency representations have also been successfully appled to optical flow problems [1] [13], using phase information as well as magnitude.

0-8186-7108-4/95 $4.00 © 1995 IEEE

3 Background

Stereo vision requires that a pair of cameras be positioned with overlapping fields of view. To simplify the forthcoming discussion we will restrict our attention to a simple case: both cameras on a horizontal plane with optical and vertical axes parallel, and known baseline.

The primary task in stereo matching is to locate pairs of pixels that are images of the same point in space. Once a correspondence has been established, it is a simple matter to determine the distance to that point using triangulation.

$$Disparity \;=\; x_{iL} - x_{iR} \;=\; \frac{Bf}{Z} \qquad (1)$$

Equation 1 describes pointwise disparity only (using the notation of Figure 3); we will show how to extend this description to surfaces in Section 4.

3.1 The Scalogram: A Unified View of Scale Space

There are many local frequency representations: spectrograms (Short Time Fourier Transforms), Wigner-Ville distributions, wavelets, and scalograms to name a few. [11] All are similar in effect, but slightly different in structure. The spectrogram uses a fixed window size at all scales and a logarithmic sampling of wavelengths. In contrast, the scalogram uses a variable window size, one which is a constant number of wavelengths long. This makes high frequencies much more localizable, and provides the necessary support for low frequencies. The scalogram is actually an instance of general wavelet functions: the scalogram comprises filter outputs from a bank of Gabor wavelets.

Figure 2 illustrates a simple signal and its scalogram, computed as follows:

$$Gabor_\lambda = \left[e^{-\left(\frac{x}{m\lambda\sigma}\right)^2} \cdot e^{-i\frac{2\pi}{\lambda}x}\right] \quad \text{for}\, x \in \left[-\frac{m\lambda}{2}, \frac{m\lambda}{2}\right]$$

$$\text{Scalogram}_R(x,y) = (R * Gabor_y)(x)$$

where R is the one dimensional input row, λ is the filter wavelength, m is the number of wavelengths to fit in the window, σ is the Gaussian parameter expressed as a fraction of the window size $m\lambda$, and $*$ denotes convolution. The signal consists of a sine wave with high frequency on the outside and one with lower frequency inside. The scalogram plots have a straightforward interpretation: the horizontal axis is the same as in the original signal (Pixel number) and the vertical axis is wavelength (in pixels). The height of the plot encodes the strength of the signal at a given location and resolution (wavelength); higher peaks mean stronger response. Phase values are most reliable on or near the magnitude peaks. The scalogram has a triangular shape because no data is plotted where the filter window would extend beyond the signal boundary.

This kind of representation is very useful for image matching. In particular, the phase measurements translate directly into disparity measurements: $disp = \frac{(\phi_L - \phi_R)}{2\pi} \cdot \lambda$. This gives us a means of generating disparities to subpixel accuracy without having to explicitly interpolate the original signal.

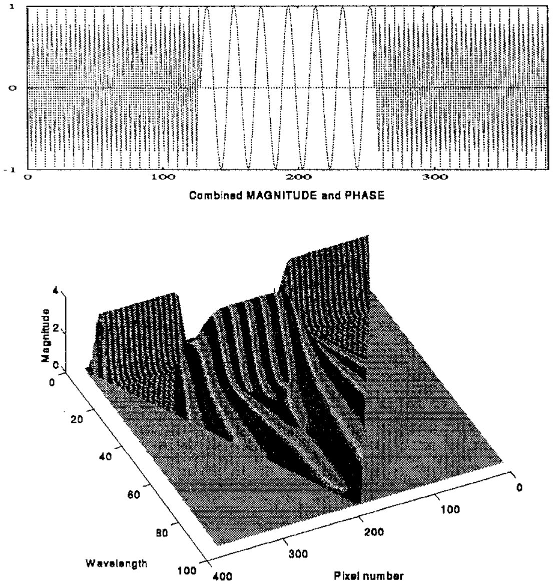


Figure 2: Double sine wave signal and associated scalogram (height is magnitude, color is phase).

3.2 Our phase method

The following foreshortening analysis applies equally well to many phase methods, but we will apply it to our phase method which uses the dense set of Gabor filters used to generate the scalogram (described in detail in [8]). Our technique is similar in spirit to that developed by Sanger [12] but is not limited to small disparities.

4 Analysis

In this section we show how perspective foreshortening is manifest in the local spatial frequency representation of stereo images.

To simplify the analysis, we assume the only object in the world is a textured flat plate that is either parallel to the image plane, or rotated about the vertical axis by some angle θ, and that the stereo cameras have parallel optical (depth) and vertical (height) axes. Thus we restrict our attention to the effects of foreshortening in one-dimensional image *scanlines*, rather than complete two-dimensional images, since all disparities will be horizontal under this assumption. Our world model will likewise be a two-dimensional slice through the three-dimensional scene. Figure 3 shows an overhead schematic of a horizontal slice through the world. By convention the parameters measuring distances in the world will be capitalized (e.g., X_S, Z_L), and those measuring pixel or camera distances will be lower case (e.g., x_{iL}, f).

Although our ultimate goal is to find the disparity between two stereo images, we must first determine

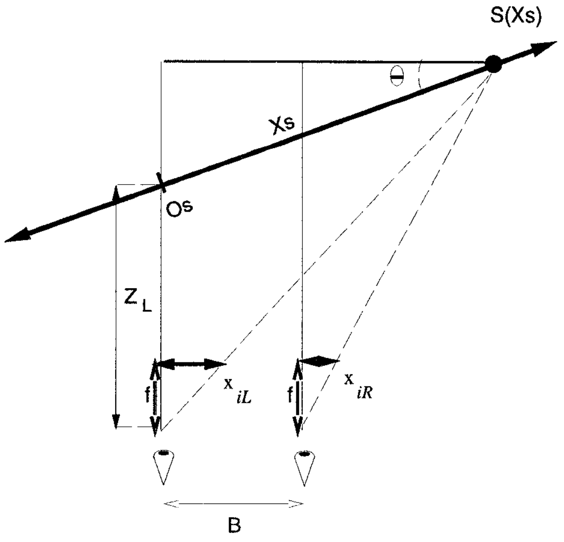


Figure 3: Overhead view of the foreshortening model. X_S is the distance from the point exactly in front of the left camera (the origin O_S at distance Z_L) to the point (S) on the plate being studied; x_{iL} and x_{iR} are the left and right pixel indices of the image of surface point S; the cameras are separated by baseline B and the surface tilts away from the cameras at angle θ.

how the appearance of the object's surface texture changes between them. This is a geometric formulation; what matters is how much of the surface is being mapped to each pixel, not the actual surface texture (i.e., color intensity). Mathematically, we want to compare the left and right sampling rates:

$$\text{Sampling ratio} \;=\; \frac{\frac{\delta X_S}{\delta x_{iL}}}{\frac{\delta X_S}{\delta x_{iR}}} = \frac{\delta x_{iR}}{\delta x_{iL}} \qquad (2)$$

Simplifying the ratio in this way, we can compute the sampling ratio (or *frequency shift*) in *image space* without having to explicitly model the distance X_S along the object. Unfortunately, this implies that we need the disparity derivative $(\delta(x_{iL} - x_{iR}))$. Since our ultimate goal is to estimate disparity, it would be best to avoid using its derivative (at best a noisy estimator) in our calculations. The remainder of this section will show how we can express this ratio with terms that do *not* require disparity derivatives.

4.1 Relating Disparity to Surface Angle

How is disparity related to the surface angle? Equation 1 gives the disparity for an individual point, but Figure 3 shows us how it varies across a surface:

$$\frac{x_{iL}}{f} = \frac{X_S \cos\theta}{Z_L + X_S \sin\theta} \qquad (3)$$

$$\frac{x_{iR}}{f} = \frac{X_S \cos\theta - B}{Z_L + X_S \sin\theta} \qquad (4)$$

Equations 3 and 4 give us expressions for x_{iL} and x_{iR} in terms of the focal length f, baseline B, distance Z_L, surface angle θ, and location on the surface X_S. Solving for X_S and setting them equal yields:

$$x_{iR} \;=\; x_{iL}\left(1 + \frac{B}{Z_L}\tan\theta\right) - \frac{Bf}{Z_L} \qquad (5)$$

And finally, recalling that disparity is the difference of the two indices:

$$\textit{disparity} \;=\; x_{iL} - x_{iR} \;=\; \frac{Bf}{Z_L} - x_{iL}\frac{B}{Z_L}\tan\theta (6)$$

Equation 6 is nearly the answer we want. It relates disparity to the scene parameters, and does not depend on knowing the actual surface location. It requires knowledge of surface distance, unfortunately, but we will eliminate this restriction below. When the surface is frontoplanar it reduces to the familiar expression relating disparity to depth from Equation 1. And for an arbitrary fixed angle θ the disparity *derivative* is constant, i.e., the disparity varies linearly with respect to the image location x_{iL}. While we won't take advantage of this property of the derivative, it could prove useful to shape-recovery techniques.

4.2 Expressing the Sampling Ratio using Image Parameters

Now that we know how the disparity and pixel locations relate to surface angle, we eliminate the derivative from Equation 2:

$$\text{Sampling ratio} \;=\; \frac{\delta\left(x_{iL}\left(1 + \frac{B}{Z_L}\tan\theta\right) - \frac{Bf}{Z_L}\right)}{\delta x_{iL}}$$

$$\textit{Geometric Form} \;=\; 1 + \frac{B}{Z_L}\tan\theta \qquad (7)$$

This expression is very interesting. It tells us that for a given flat surface, the sampling ratio is *constant* over both images of the surface. In other words, the local spatial frequencies of the left and right images are related by a simple constant scale factor. However, Equation 7 would be useless in a stereo matcher since it requires knowledge of the depth Z_L. A program that computed depth given depth would not be very impressive. To eliminate Z_L we solve Equation 6 for $\frac{B}{Z_L}$ and replace it in Equation 7, giving us this final expression for Frequency Shift (aka Sampling Ratio):

$$\textit{Projected form} \;=\; 1 + \frac{\textit{disparity}\tan\theta}{f - x_{iL}\tan\theta} \qquad (8)$$

This is what we want! Equation 8 relates parameters in the image plane to the surface slope θ, but requires neither prior knowledge of the distance to the object nor an estimate of the disparity derivative. We will see how to manage its parameters algorithmically in Section 5.

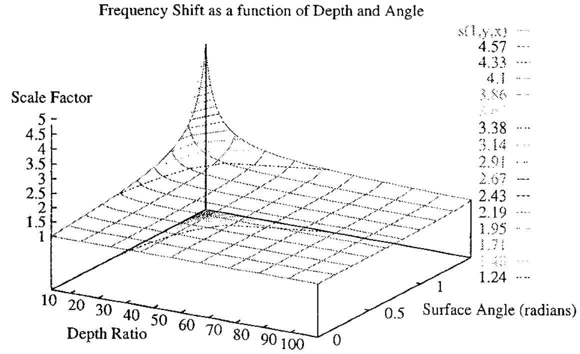


Figure 4: Frequency Shift as a function of Depth and Angle. Depth is unitless relative to the baseline, and varies from 3 to 100. Angle varies from zero to 85°.

4.3 Applicability

How important is this foreshortening analysis? More specifically, how often do situations arise in which the assumption that a surface is frontoplanar causes problems for stereo systems?

To show this we will use the geometric formulation of the Scale Factor from Equation 7. For this analysis we consider the ratio of depth over baseline $\frac{Z_L}{B}$ to be a single variable, unitless *depth* (relative to the camera baseline). For example, the distance between a person's eyes would be 1, the distance to their computer monitor 4-6, and the distance to the far wall in a typically small three-person graduate student office about 100. Figure 4 plots the near-complete Scale Factor space for a person looking at objects in such an office.

Suppose we assume that surface depth and orientation are uniformly distributed throughout a scene. Then we can compute the probability that a surface will require at least a 10% correction term by finding the area under the 1.1 Scale Factor contour curve. The derivation is given in [8], but the result is that given a uniform distribution of angles and depths, the probability that a surface will require at least a 10% correction is 0.210355. Try it out; if you're sitting in an office, see if you can find one sharply foreshortened surface for each set of four nearly head-on surfaces in your immediate vicinity.

Of course the probability of finding foreshortened surfaces depends very much on the domain. Robot vehicles like Carnegie Mellon's NAVLAB often use a very wide baseline, on the order of one meter. With the nearest visible ground point about five meters away, depth ratios of 5 to 20 are common in this domain. In that range, the probability of finding a foreshortened surface jumps to better than one in three (see Table 1). Inspection robots typically use much smaller baselines, but the probability of foreshortening is still significant (nearly one in twelve). These results suggest that many stereo vision systems could benefit from an analysis that considers the effects of foreshortening.

Depth Range	P(10% effect)	Example Domain
0-100	0.210355	*Human in office*
5-20	0.354404	*Robot Vehicle*
30-100	0.0808227	*Inspection Robot*

Table 1: Probability that a surface exhibits 10% variation between images due to perspective foreshortening. The distribution of surfaces is assumed to be uniform within the range of orientation angles from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, and given depth ratios (distance divided by baseline).

5 Application

The analysis in Section 4 is not only theoretically interesting, it can also improve the performance of real stereo algorithms such as [3], [12] and our method [8].

5.1 Extending Phase-based Stereo Algorithms

Some have argued that a small number of Gabor filters are sufficient for stereo matching. [3] The idea is that although the phase may vary slightly across nearby frequencies, the variation is small enough that the error introduced in measuring it at what might be the wrong frequency will be insignificant. But we have seen that frequency shifts of even 10% can occur often. Instead of introducing error by sampling at the wrong frequency, we can turn these perturbations to our advantage by using them to confirm hypotheses of surface slant.

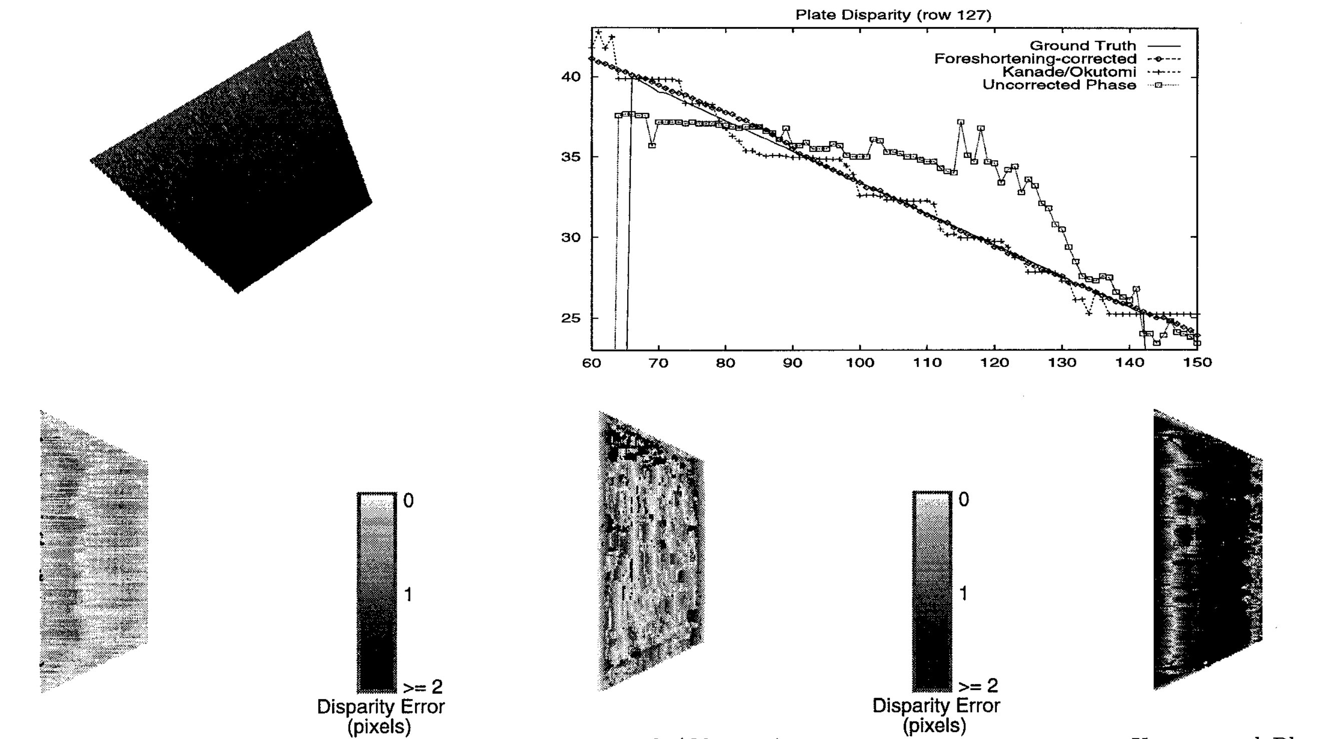
We will need a dense sampling of the phase space to get the most accurate results. We will also interpolate phase values between adjacent frequencies when possible. The image scalogram provides a useful framework for such computations, and will be used as the basis for our foreshortening-corrected stereo algorithm.

Our method outlined in [8] uses a global minimization strategy to find the best disparity from a list of candidates. This framework makes it easy to include a foreshortening correction term: in addition to searching disparity space, we also search over surface angle. Pseudocode for this revised algorithm is given in Table 2. The only difference between this and the original algorithm is the presence of the correction term on the right image phase measurements.

5.2 Results

Consider the stereo pair in Figure 1. It shows a synthetic stereo image pair of a flat plate rotated 65 degrees from the image planes, with the image of a city scene texture-mapped onto the plate. The actual disparity map (known from the 3D world model) and differences between the ground truth and disparity computed by three stereo methods are presented in Figure 5.

For this demonstration of the foreshortening-corrected algorithm, a set of 501 potential disparities were considered (0 to 50 in steps of 0.1), and the angle was fixed at 65 degrees. The RMS error of this result was 0.38 pixels over the entire plate, with $\sigma = 0.63$. The bulk of this error can be attributed to two causes: the dark spots and a subtle systematic error over the



Foreshortening-corrected Kanade/Okutomi Uncorrected Phase

Figure 5: Ground Truth and computed disparity maps for a surface angled at 65°. The top row shows ground truth in perspective on the left, a graph of a representative scanline from all methods on the right. The bottom row shows differences between actual disparities and those computed by the foreshortening-corrected method, Kanade/Okutomi and the uncorrected phase method, for pixels that image the plate; darker values denote larger errors. Only differences between 0 and 2 pixels are shown, errors larger than 2 pixels appear as a 2 pixel error.

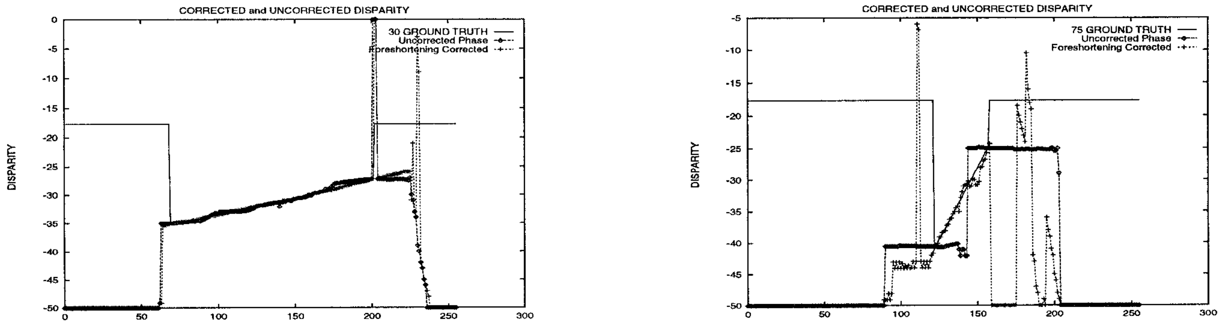


Figure 6: Ground truth and disparity (computed by both the uncorrected and foreshortening-corrected phase methods) for the center scanline of the city scene at 30 and 75 degrees.

```
Given: A pair of greyscale images, lists of potential disparities and surface angles, focal length $f$.

For each row
    Compute Left and Right Scalograms $L$ and $R$
    For each column $c$
        For each disparity $d$
            For each angle $a$
                $correction = 1 + \frac{d \tan a}{c \tan a - f}$
                $error = \sum_{\lambda : \rho(\lambda) > threshold} \rho_L(c, \lambda) \cdot$
                    $|\Delta\phi_{ideal}(d, \lambda) - (\phi_L(c, \lambda) -$
                        $\phi_R(c + d, \lambda \cdot correction))|_{2\pi}$
        Return $d$ (and $a$) that yield minimum $error$
```

Table 2: Pseudocode for the foreshortening-corrected algorithm. Column index c is zero in the center of the image.

surface. The spots most likely arise from an artifact of the rendering process which caused a few nearby pixels in one image to map to the same intensity. The more subtle effect is that the disparity error, while within measurement bounds at the ends and center of the plate, varies by as much as 0.5 pixels between the center and end of the plate (see Figure 5, upper right).

The Kanade-Okutomi variable-window refinement method [6] uses a statistical analysis to grow the window from 3x3 to some maximum, stopping when an error criterion (based on local changes in intensity and disparity) is exceeded. For this test we let disparity vary between 0 and 50 pixels (as in our method), let the window size vary from 3 to 21 pixels, and ran for 10 iterations. It approximated the surface shape well, but produced many more outliers and quantized the flat tilted surface into several stair-step frontoplanar patches (see Figure 5, upper right). The RMS error of this method was 0.99 pixels over the entire plate, with $\sigma = 2.36$.

The uncorrected phase method results are also shown in Figure 5. The same 501 potential disparities were considered, but foreshortening correction was not applied. The RMS error of this result was 3.77 pixels over the plate, with $\sigma = 6.23$. The main source of error was a general flattening of the entire plate.

Other Rotation Angles The uncorrected method does reasonably well with small angles, but at slants greater than 30° its performance degrades by several pixels [8] (see Figure 6). In contrast, the foreshortening-corrected method performs well even at 75°, though at 80° the systematic error becomes more apparent.

6 Acknowledgements

Thanks to Paul Heckbert for many useful discussions, and to Larry Matthies for providing the city scene used in the synthetic examples.

References

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[2] Peter N. Belhumeur. A binocular stereo algorithm for reconstructing sloping, creased, and broken surfaces in the presence of half-occlusion. In *International Conference on Computer Vision*, pages 431–438, 1993.

[3] David J Fleet, Allan D Jepson, and Michael R M Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, March 1991.

[4] Michael R. M. Jenkin and Allan D. Jepson. Recovering local surface structure through local phase difference measurements. *CVGIP: Image Understanding*, 59(1):72–93, January 1994.

[5] David G. Jones and Jitendra Malik. Determining three-dimensional shape from orientation and spatial frequency disparities part ii - using corresponding image patches. Technical Report UCB/CSD 91/657, University of California, Berkeley, Computer Science, October 1991.

[6] Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. In *DARPA Image Understanding Workshop Proceedings*, pages 383–398, September 1990.

[7] John Krumm. Shape from texture and segmentation using local spatial frequency. Technical Report CMU-RI-TR-93-32, Carnegie Mellon Univiersity Robotics Institute, May 1994.

[8] Mark W. Maimone and Steven A. Shafer. Modeling foreshortening in stereo vision using local spatial frequency. Technical Report CMU-CS-95-104, Carnegie Mellon Univiersity Computer Science Department, January 1995.

[9] Jitendra Malik and Pietro Perona. A computational model of texture segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–332, June 1989.

[10] Dale J. Panton. A flexible approach to digital stereo mapping. *Photogrammetric Engineering and Remote Sensing*, 44(12):1499–1512, December 1978.

[11] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, October 1991.

[12] Terence D Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418, 1988.

[13] Yalin Xiong and Steven A. Shafer. Hypergeometric filters for optical flow and affine matching. In *International Conference on Computer Vision*, 1995.