In 1998 IEEE
International Conference on Robotics and Automation
Leuven, Belgium - May 1998

Stereo Vision Enhancements for Low-Cost Outdoor Autonomous Vehicles

Alonzo Kelly and Anthony Stentz

Field Robotics Center
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890
email: alonzo@ri.cmu.edu, url: http://www.frc.ri.cmu.edu/~alonzo
email: axs@ri.cmu.edu, url: http://www.frc.ri.cmu.edu/~axs

Abstract

The contemporary implementation of software systems for off-road navigation on conventional computing involves a continuous struggle to develop more computationally efficient algorithms. A basic trade-off exists between system performance and reliability for any given level of efficiency of the software. Hence, the only way to improve both performance and reliability is to improve the computational efficiency of the underlying algorithms. Dense stereo vision algorithms can easily exhaust almost all available computer cycles and are therefore prime candidates for the development of more efficient approaches. This paper presents some techniques that can be used to improve the efficiency and/or reliability of dense stereo for off road autonomous vehicles.

1 Introduction

The need for high throughput perception algorithms has been acknowledged for some time [2][4][13][12] in the field of autonomous vehicle navigation. In our earlier work [7], the authors have presented several adaptive techniques for the implementation of active stereo vision and terrain mapping in off-road scenarios. These techniques included:

- software modulation of the distance beyond which geometry is processed to correspond to the instantaneous vehicle response distance.
- software modulation of the width of the window of ranges processed to correspond to the distance moved by the vehicle in each perceptual cycle.
- software modulation of the range pixel aspect ratio to compensate for the average expected elongation of pixels when projected onto the groundplane.

This paper builds on this earlier work and introduces several new techniques to further increase performance. To the degree that the techniques increase performance, they enable the use of lower cost computing and sensing systems for a fixed performance requirement and contribute to the development of lower cost fieldworthy autonomous vehicles.

Computer stereo vision is an area that has been studied for at least two decades [11]. Our work in the area reuses ideas of hierarchical processing of images [1][4] and the use of relaxation techniques that reflect the interdependencies of various parts of the problem [9][5]. More recently, researchers have recognized the importance of adaptive signal matching techniques [14]. Our work is in the spirit of this trend while also introducing a form of relaxation and applying it to an application where the disparity gradient is usually high.

2 Preliminaries

This section introduces terminology and the core concepts upon which the paper is based.

2.1 Coordinate Conventions

The angular coordinates of a pixel will be expressed in terms of horizontal angle or **azimuth** ψ, and vertical angle or **elevation** θ. Three orthogonal axes are considered to be oriented along the vehicle body axes of symmetry. Generally, we will arbitrarily choose z up, y forward, and x to the right:

- x - **crossrange**, in the groundplane, normal to the direction of travel.
- y - **downrange**, in the groundplane, along the direction of travel.
- z - **vertical**, normal to the groundplane.

Certain vehicle dimensions that will be generally important in the analysis are summarized in the following figure. One distinguished point on the vehicle body will be designated the vehicle control point. The position of this point and the orientation of the associated coordinate system is used to designate the pose of the vehicle.

The wheelbase is L, and the wheel radius is r. The height of the sensor above the groundplane is designated h and its offset rear of the vehicle nose is p. The height of the undercarriage above the groundplane is c. Range measured from the sensor is designated R.

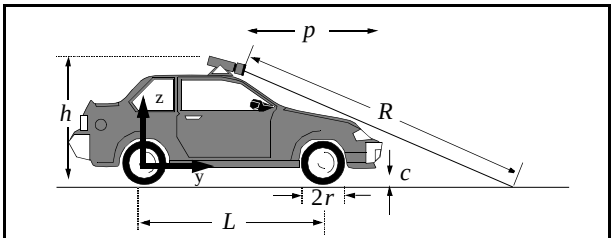

Figure 1: Important Dimensions

We will sometimes distinguish range, R measured in 3D from a range sensor, and the projection of range Y onto the groundplane. Generally, both will be measured forward from the sensor unless otherwise noted. The velocity of the vehicle will be denoted V.

We will describe the coordinates of a point both with respect to an image and with respect to a correlation window within the

image. The figure introduces the conventions used in the paper.

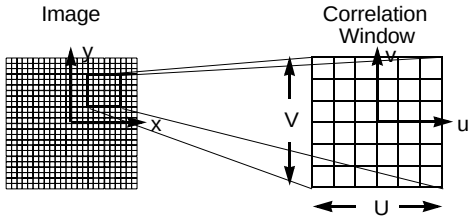

Figure 2: Image and Correlation Window Coordinates. A local coordinate system (x,y) is attached to the image plane at the central pixel of the reference image and another (u,v) is attached to the correlation window.

2.2 Normalized Disparity

We will present our work in the context of binocular (two-eyed) stereo, though none of our fundamental results depend on the number of cameras used. The basic binocular stereo ranging rela-

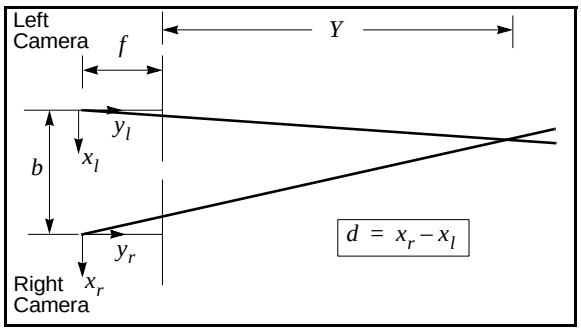

Figure 3: Basic Triangulation in Binocular Stereo. The baseline is the perpendicular distance between the optical axes. The range is measured normal to the reference image plane. The disparity is the difference in image coordinates of corresponding points measured along the epipolar line. It also depends on the focal length.

tionship for perfectly aligned cameras is derived from similar triangles. It relates disparity d, range Y, baseline b, and focal length f:

$$Y = bf/d$$

It is useful to remove the dependence of disparity on the focal length by expressing disparity as an angle. Define the **normalized disparity** thus:

$$\delta = \frac{d}{f} = \frac{b}{Y}$$

2.3 Disparity Gradient

The **disparity gradient** is the spatial derivative of disparity in some coordinate system. Under the assumption that the images have been rectified into perfect epipolar geometry, the disparity is a scalar field over the image plane known also as the **disparity image**. An associated vector field, the disparity gradient image, can be derived from it:

$$\vec{\nabla}\delta(x,y) = \frac{\partial}{\partial x}\delta(x,y)\hat{i} + \frac{\partial}{\partial y}\delta(x,y)\hat{j}$$

3 Geometric Decorrelation

It is well known that the gradient of range measured from the image plane of the reference camera introduces a disparity gradient across the correlation window in area-based stereo [3]. This section investigates the qualitative behavior of the disparity gradient for horizontal and vertical baseline stereo systems.

3.1 Geometric Decorrelation for Horizontal and Vertical Baselines

Let the reference image be defined as the image whose coordinate system is used to express the resulting range image. Without loss of generality, we will take the left image to be the reference image for horizontal baselines and the top image will play this role for vertical baselines.

The following figures show that a fixed angular region projected from different positions will image different areas on the groundplane.

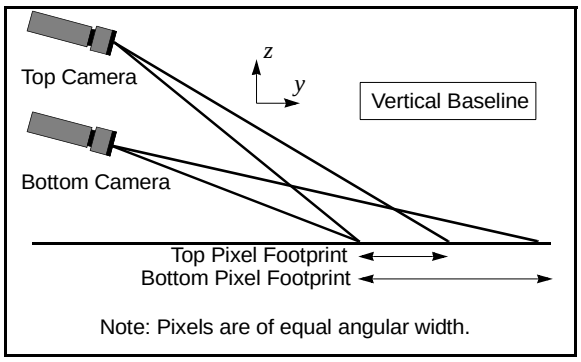

Figure 4: Pixel Foreshortening - Vertical Baseline. A fixed angular region projected from different positions images different sized areas on the groundplane.

This is true for both horizontal and vertical baselines as shown in the next figure.

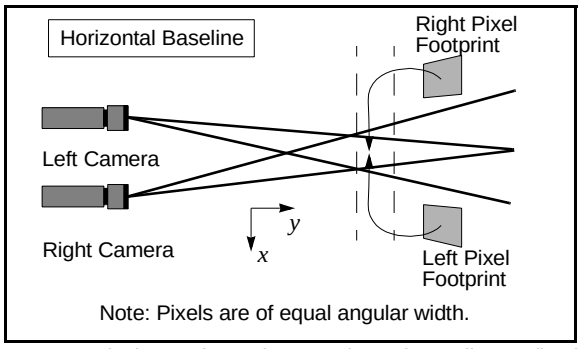

Figure 5: Pixel Foreshortening - Horizontal Baseline. A fixed angular region projected from different positions images different sized areas on the groundplane.

From the reverse perspective, for flat terrain, a region on the groundplane which projects onto a regular rectangle in the reference image will project onto a region of different size and/or

shape in the other image as shown below.

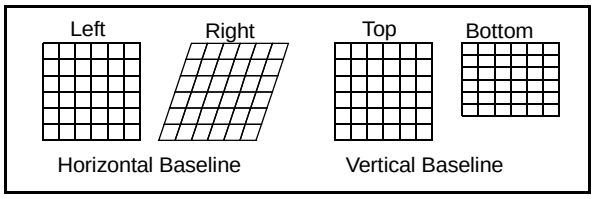

Figure 6: Geometric Decorrelation. A correlation window in the reference image will have its corresponding window distorted in the other image as shown.

This foreshortening effect goes by several names in the literature. Its effect is to cause attempts to correlate corresponding regions of images to fail because the right intensity values are in the wrong places. No rigid transformation of the reference image correlation window results in the other image correlation window.

We will call the effect **geometric decorrelation**. All approaches to stereo have different methods which attempt to deal with it. The rest of the paper introduces two techniques for dealing with this problem in increasing order of sophistication.

3.2 Relationship to Disparity Field

Generally, of course, the correct mapping from a region in the reference image to the other image is the very thing we are looking for in stereo - the disparity image. If x and y denote coordinates in the image plane, then disparity $d(x, y)$ in an ideally rectified, left-referenced, horizontal baseline system is determined by the following equations relating the coordinates of corresponding pixels in each image.

$$x_R = x_L + d(x_L, y_L) \qquad y_R = y_L$$

Given this relationship, it is clearly possible to compute the mapping of any region from the left image to its corresponding region in the right. When disparity is continuous, every pixel's surrounding region in the reference image maps onto a corresponding region in the other image. When disparity is not continuous, it maps onto a set of regions. Some cases are indicated in the following figure for several possibilities for the geometry of the surface being imaged.

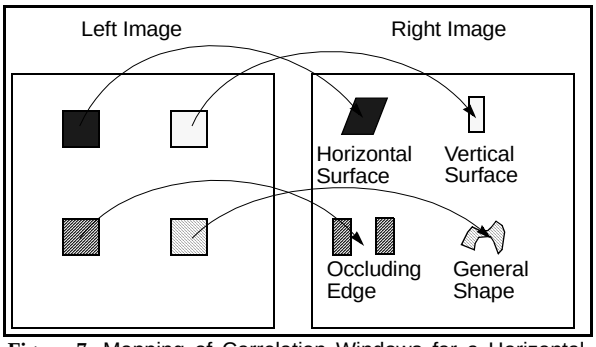

Figure 7: Mapping of Correlation Windows for a Horizontal Baseline. By definition, the disparity image $d(x, y)$ maps pixels and hence regions from the reference image to the other image. Some typical cases are indicated.

3.3 Fundamental Inconsistency of Area-Based Stereo

The observation that geometric decorrelation occurs leads to the observation that traditional or "area based" stereo incorporates a fundamental inconsistency, because:

- Disparity cannot be everywhere different **and** have zero gradient.
- Disparity can't be both uncorrelated (everywhere different) and correlated (constant in a small window).

This inconsistency arises as soon as we attempt to match two correlation windows of the same shape, any fixed shape, or any fixed relationship - one of each comes from each image. Unless the disparity is constant across the window, corresponding identically shaped "regions" can never correspond perfectly - only points can.

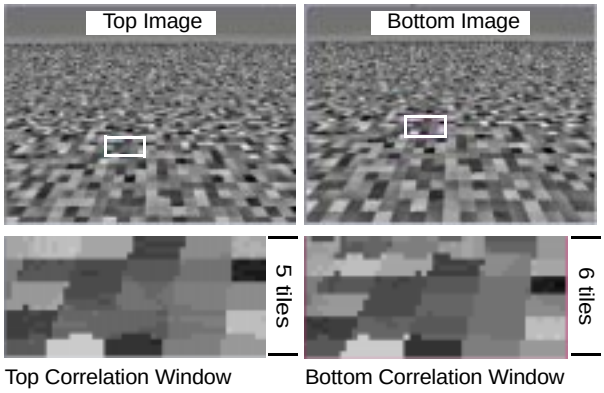

Figure 8: Example of Geometric Decorrelation. Corresponding correlation windows for two synthetic images of perfect geometry. The last row of each corresponds, but the bottom window has an extra row at the top of the window.

3.4 Results

Figure 8 illustrates geometric decorrelation for a perfectly aligned vertical pair for cameras looking at a flat surface, which has randomly generated greyscale tiles. The images were produced by the ray tracer of our simulator so their geometry is perfect. Note that although the bottom rows of the correlation windows correspond, the bottom correlation window has an extra row at its top. The balance of the paper will introduce techniques for dealing with this which each have varying degrees of sophistication and success.

4 Modulation of Correlation Window Aspect Ratio

In this section, we derive the optimal fixed shape of a stereo correlation window for flat terrain from the perspective of minimizing the effects of geometric decorrelation.

4.1 Disparity Gradient of Flat Terrain

A simple expression, accurate to first order, is available for the gradient of disparity in an image of flat terrain. In such an image, the gradient is wholly vertical. It is related to the Δy spanned by the correlation window height $\Delta\theta$. From earlier analysis [8] for flat terrain we know that a small change in image elevation angle moves a range pixel a corresponding distance along flat terrain given by:

$$\Delta y = \frac{y^2}{h}\Delta\theta$$

Differentiating the earlier expression for normalized disparity with respect to range leads to:

$$\Delta\delta = -\frac{b}{y^2}\Delta y$$

Substituting the first relationship into the second shows that the disparity gradient is equal to the **normalized baseline** - the ratio of baseline to sensor height [10]. Notice that two quadratic rela-

$$\frac{\Delta\delta}{\Delta\theta} = -\frac{b}{h}$$

tionships have cancelled to leave us with the convenient rule that linear variation in range leads to a linear variation in disparity.

Of course, the disparity gradient at any point in an image is a direct function of the real range gradient of rough terrain. In general, it may vary significantly as the terrain slope varies, but the above figure is an acceptable approximation for many of our purposes.

4.2 Coefficient of Geometric Decorrelation for Uniform Disparity Gradient

The distortion of correlation windows is purely a geometric matter - independent of the image data itself and dependent solely on the disparity gradient field. We can easily derive an expression for the total displacement from their nominal positions of all pixels in the window.

Recalling Figure 2, let us attach a local (u, v) coordinate system to a window of size (U, V) in some arbitrary image plane linear units. Let the disparity gradient be given by:

$$\vec{\nabla}\delta = \left[\frac{\partial\delta}{\partial u}\ \frac{\partial\delta}{\partial v}\right]^T = \left[\delta_u\ \delta_v\right]^T$$

and *assume it is constant over the correlation window.* Then the **coefficient of geometric decorrelation** C_{err} will be defined as the weighted integral of the area of the window where the weight of each differential region is given by its shift relative to its nominal undistorted position. Thus, if $\vec{\rho}$ is the position vector in the image plane, then:

$$C_{err} = \int_{-\frac{U}{2}}^{\frac{U}{2}}\int_{-\frac{V}{2}}^{\frac{V}{2}} \vec{\nabla}\delta \bullet \vec{\rho}\, du dv = \int_{-\frac{U}{2}}^{\frac{U}{2}}\int_{-\frac{V}{2}}^{\frac{V}{2}} (\delta_u u + \delta_v v)\, du dv$$

Therefore, when the disparity gradient is constant:

$$C_{err} = \frac{\delta_u U^2 V}{4} + \frac{\delta_v V^2 U}{4}$$

Notice that the decorrelation depends only on the disparity gradient and the window dimensions.

4.3 Optimal Shape of the Correlation Window

The above expression can be optimized if another constraint equation is available. Let the area of the correlation window remain constant while the width and height vary dependently. Then, setting the derivative to zero, there results:

$$\frac{U}{V} = \frac{\delta_v}{\delta_u}$$

Thus *the optimum aspect ratio of the correlation window is the inverse of the ratio of the corresponding components of the disparity gradient.* This is intuitively appealing because it requires that the extent of the window be minimized in the direction where disparity changes most. This result explains why outdoor stereo systems tend to perform better when the correlation window is wider than it is high. On average, the disparity gradient is mostly vertical in the image plane in outdoor settings.

4.4 Results

We have begun to use this rule in our work in order to minimize the effects of geometric decorrelation. Figure 9 clearly shows a dramatic improvement in the accuracy of disparity images of flat terrain based on simply changing the aspect ratio of the correlation window.

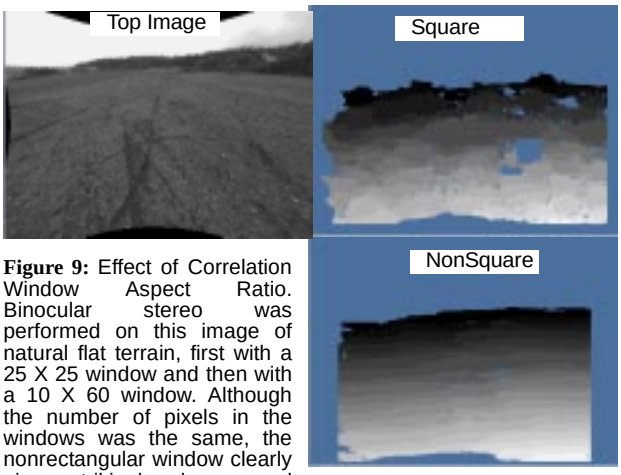

Figure 9: Effect of Correlation Window Aspect Ratio. Binocular stereo was performed on this image of natural flat terrain, first with a 25 X 25 window and then with a 10 X 60 window. Although the number of pixels in the windows was the same, the nonrectangular window clearly gives strikingly cleaner and more accurate disparity images.

Later sections provide better mechanisms which achieve the same types of improvements.

5 Constant Modulation of Correlation Window Shape.

In the last section, we varied the aspect ratio of the correlation window in order to minimize the effects of geometric decorrelation. The window area, however, remained constant as did its shape. This section introduces techniques that modify the size and/or shape of the correlation window based upon expectations.

It may seem to the reader that the introduction of expectations would make stereo more error prone when those expectations are violated. However, note that the traditional correlation of rectangular subwindows *amounts to an assumption that the disparity gradient is zero.* This is equivalent to the assumption that the terrain is nearly **vertical**.

Hence, traditional stereo already incorporates expectations and the particular ones used are not the ones that would be chosen in off-road autonomous vehicles which tend to operate on nearly horizontal terrain most of the time[1]. We have already calculated that the disparity gradient is equal to the normalized disparity in this case.

One simple way to introduce expectations is to introduce a warp-

1. There is the argument that vertical surfaces often constitute obstacles and that stereo should be tuned to detect them. We accept that argument while striving to develop stereo algorithms which need no such tuning. Our navigator also avoids unknown terrain which would be generated by poor ranging to vertical surfaces so this point is less relevant to us.

ing function defined on the correlation window of the form

$$D_{warp}(u, v) = \int_{l} \vec{\nabla}\delta(u, v) \bullet \vec{ds}$$

This function is the integral of the assumed disparity gradient $\nabla\delta(u, v)$ along the epipolar line l. It therefore gives the displacement of a pixel along the epipolar line from its nominal position with respect to the window center if the disparity gradient were zero. It encodes the change in shape of a rectangular correlation window when mapped to the other image.

Computation of disparity is accomplished through maximizing some measure of window similarity such as normalized correlation by searching a sequence of proposed or candidate disparities.

Consider the horizontal baseline case. If $L(x, y)$ is the left normalized image and $R(x, y)$ is the right normalized image, then the normalized correlation computed for each pixel as a function of a proposed disparity d is computed from the double integral:

$$C(x, y, d) = \frac{1}{UV} \int_{-\frac{U}{2}}^{\frac{U}{2}} \int_{-\frac{V}{2}}^{\frac{V}{2}} L(x + u, y + v) R(x + u - d, y + v) du dv$$

In order to use the window warping function proposed above we replace the constant proposed disparity d above by the **offset warping function**:

$$\Delta(u, v) = d + D_{warp}(u, v)$$

where the first term represents the candidate relative position of the window and the second encodes its shape. Note that the window origin is always unwarped:

$$D_{warp}(0, 0) = 0$$

This approach corresponds to simply warping the window and then correlating it in candidate positions along the epipolar line as usual as illustrated below.

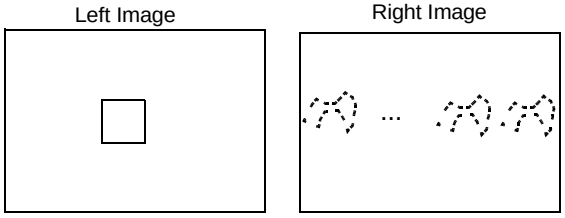

Figure 10: Correlation with Window Warping. We can warp the left correlation window by an arbitrary function and then search for its position in the right image.

In practice, the warping function can be computed by assuming a constant for the disparity gradient based on the results of earlier sections and then integrating it. Some examples were given in Figure 7.

5.1 Results

Figure 11 illustrates the use of a constant warping function on simulated flat terrain. Because synthetic images were used,

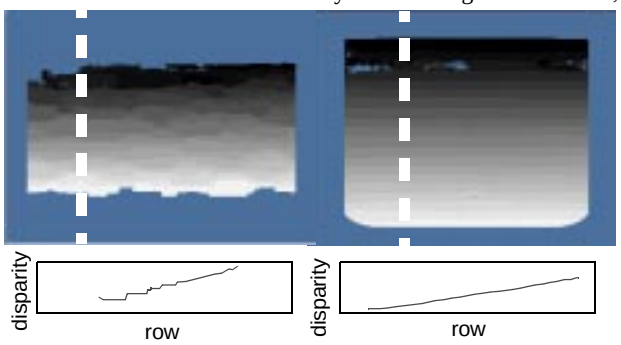

Figure 11: Constant Warping Function. The constant warping function technique was applied to the synthetic stereo pair of Figure 6. The left disparity image is based on no warping and the right is based on using the normalized baseline warping function.

ground truth is a perfectly flat patch of terrain and a virtually linear disparity image. The disparity versus row curves for the right image are clearly smoother and more accurate whereas the left image shows the disparity steps that typically occur when the gradient is high.

6 Disparity Relaxation Stereo

The reader may have noticed that since the proposed approach assumes a value for the disparity gradient (and hence disparity, to within a constant) from which a new disparity is computed, the problem is somewhat circular. This circularity leads to a formulation of stereo as a relaxation of the disparity image. In each iteration of the algorithm, we assume that the current disparity image is locally correct (i.e. of the right shape) but that it may need to have its position adjusted somewhat. This section derives such an algorithm.

6.1 Principle

Let $d^k(x, y)$ denote the disparity image at iteration k, and let d denote the candidate disparity being searched in the evaluation of the 3D scalar field $C(x, y, d)$. Unlike in the previous section, we allow the warping function, now denoted as $D^k_{warp}(x, y, u, v)$, to vary with position in the image. However, we will suppress the dependence on position in the notation for readability.

We compute the warping function from the initial estimate of disparity but continue to search for the position of the best correlation. Thus:

$$D^k_{warp}(u, v) = \{d^k(u, v) - d^k(0, 0)\}$$

where every term above is also a function of image position. This gives the warping function offset by a candidate disparity as:

$$\Delta^{k+1}(u, v) = d + \{d^k(u, v) - d^k(0, 0)\}$$

We can now define a quantity called **relative disparity** as the difference between the candidate disparity and the current disparity of a pixel, thus:

$$d^k_{rel} = d - d^k(0, 0)$$

Reorganizing terms leads to an expression for the offset warping

function in terms of the relative disparity and the original disparity image:

$$\Delta^{k+1}(u,v) = d^{k}_{rel} + d^{k}(u,v)$$

Thus, the warped position of a pixel is given by the position in the reference image, plus the current estimate of disparity plus the current candidate relative disparity. The relative disparity would typically be in the range of small signed integers such as:

$$-n \le d^{k}_{rel} \le n$$

The process is, for each pixel position (x, y), move it to the other image, offset it by its current estimated disparity and then search, using it and its appropriately transformed neighbors, for a better match on either side of the current estimate.

For iteration k+1, the correlations are computed as:

$$C^{k+1}(x,y,d_{rel}) = \frac{1}{UV}\int_{-\frac{U}{2}}^{\frac{U}{2}}\int_{-\frac{V}{2}}^{\frac{V}{2}} L(x+u,y+v)R(x+u-\Delta^{k+1}(u,v),y+v)\,du\,dv$$

6.2 Refinements

Further refinements are possible which incorporate and/or modify the ideas of other researchers for adaptive signal matching. First, as the iteration number grows, the disparity image becomes more accurate and less search is required. Hence, the disparity window searched can be quickly reduced to increase speed.

Second, there is of course a trade-off between signal to noise and distortion where larger correlation windows have better signal to noise ratios in the absence of distortion while being more susceptible to errors due to distortion. Also, large disparity search regions increase the likelihood of false matches due to nearly repetitive texture. However, as the iteration number grows, the locally correct shape of the disparity image implies that distortion decreases and that the disparity search region can be reduced. We have found it possible and profitable to reduce the correlation window to a very small size as the relaxation proceeds.

6.3 Results

We have used the idea of disparity relaxation based stereo in its simplest form. A single iteration is performed where the disparity image of flat terrain forms the initial estimate. This estimate can be computed from the reference camera field of view and resolution and its position and orientation with respect to the flat terrain.

Although outdoor terrain is certainly not always flat, it is often closer to being flat than it is to being vertical as our results in Figure 12 show. Traditional stereo applied to this pair generates uneven artifacts and provides no data in the safest places to drive. Use of a nonsquare window improves matters. Using relaxation - even on a flat initial estimate - generates a dense disparity image of relatively high quality which correctly represents the sides of the ravine as well as its bottom.

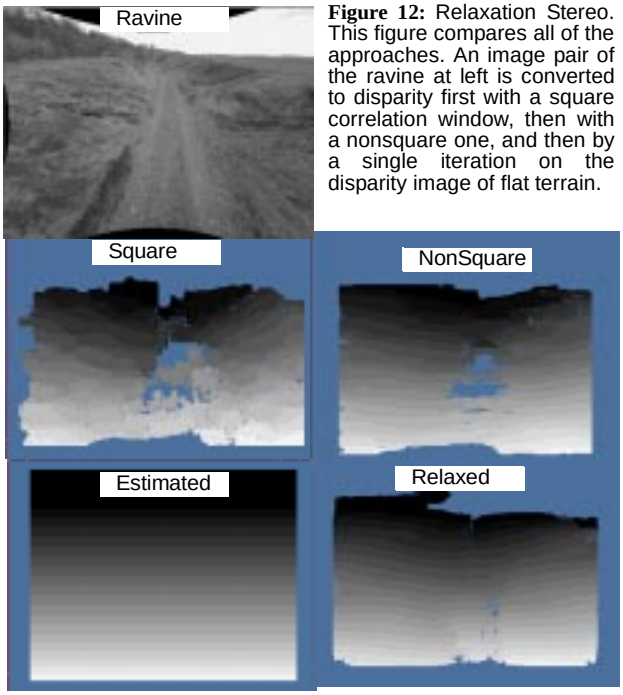

Figure 12: Relaxation Stereo. This figure compares all of the approaches. An image pair of the ravine at left is converted to disparity first with a square correlation window, then with a nonsquare one, and then by a single iteration on the disparity image of flat terrain.

7 Conclusions

We have presented a new approach to stereo vision that has advantages in situations where some amount of expectation of the shape of the environment can be employed. Although its potential applicability may be broader, we have applied it initially to outdoor autonomous vehicles. In this domain, at least where vehicles are tested today, the terrain surface is:

- almost never normal to the optical axis of the reference camera
- usually tilted away from the camera
- mostly smooth, but
- punctuated by occluding edges

Our approach seems to improve many important characteristics of stereo systems:

- unlike in more traditional approaches, flat terrain with high disparity gradient is correctly computed as flat.
- accounting for geometric decorrelation improves robustness in the presence of noise, distortion, and nearly repetitive texture.
- reduced search windows improves speed and immunity to nearly repetitive texture.

Disparity relaxation both improves speed and reduces memory

requirements. In the figure below, the cube represents the 3D

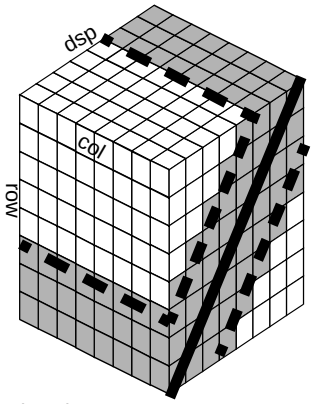

Figure 13: Speed and Memory Improvements. use of disparity relaxation improves speed and reduces required memory.

array of correlation scores $C(x, y, d)$. The thick diagonal line represents the initial estimated disparity image. Searching a few disparities on either side of the estimate implies that the white cells need never be computed or stored in memory. By using the relative disparity as the array index, a rectangular array can still be used.

8 Acknowledgments

This research was sponsored by ARPA under contracts “Perception for Outdoor Navigation” (contract number DACA76-89-C0014, monitored by the US Army Topographic Engineering Center) and “Unmanned Ground Vehicle System” (contract number DAAE07-90-C-R059, monitored by TACOM).

9 References

[1] S. T. Barnard, “Stochastic Stereo Matching over Scale”, *International Journal of Computer Vision*, pp 17-32, Jan 1985.

[2] M. Daily et al., “Autonomous Cross Country Navigation with the ALV”, In Proc. of the 1988 IEEE International Conference on Robotics and Automation, Philadelphia, Pa, April 1988; pp. 718-726.

[3] O. Faugeras, *Three Dimensional Computer Vision: a Geometric Viewpoint*. MIT Press, 1993.

[4] H. P. Moravec, “Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover”, Ph. D. Thesis, Stanford University, 1980.

[5] M. J. Hannah, “Bootstrap Stereo”, in Proc. ARPA Image Understanding Workshop, College Park, MD, Apr 1980, pp 201-208.

[6] A. J. Kelly, “An Intelligent Predictive Control Approach to the High-Speed, Cross Country Autonomous Navigation Problem”, Ph. D. thesis, Robotics Institute, Carnegie Mellon University, June 1995.

[7] A. Kelly, and A. Stentz, "Minimum Throughput Adaptive Perception for High Speed Mobility", IROS, Grenoble, France, 1997.

[8] A. Kelly, and A. Stentz, “Rough Terrain Autonomous Mobility, Part 1: A Theoretical Analysis of Requirements”, Autonomous Robots, 4(2) 1998, Kluwer Academic.

[9] Y. C. Kim and J. K. Aggarwal, “Positioning 3D Objects Using Stereo Images”, *IEEE Journal of Robotics and Automation*, vol RA-3, no. 4, pp. 361-373, Aug 1987.

[10] H. K. Nishihara et. al., “ARPA Unmanned Ground Vehicle Stereo Vision Program, Final Report Dec 1991 to Dec 1993”, Teleos Research.

[11] D. Marr and D. Poggio, “A Computational Theory of Human Stereo Vision”, Proceedings of the Royal Society of London, vol B204, pp. 301-328, 1979.

[12] L. Matthies, “Stereo Vision for Planetary Rovers”, *International Journal of Computer Vision*, 8:1, 71-91, 1992.

[13] K. E. Olin and David Tseng. “Autonomous Cross Country Navigation”, IEEE Expert, August 1991.

[14] M. Okutomi and T. Kanade, “A Locally Adaptive Window for Signal Matching”, Proceedings of the Third international Conference on Computer Vision” Osaka, Japan, Dec 4-7, 1990.