Pyramid Based Depth from Focus

Trevor Darrell
Kwangyoen Wohn

Grasp Laboratory
Computer and Information Science
University of Pennsylvania
trevor@grasp.cis.upenn.edu
wohn@grasp.cis.upenn.edu

ABSTRACT

A method is presented for depth recovery through the analysis of scene sharpness across changing focus position. Modeling a defocused image as the application of a low pass-filter on a properly focused image of the same scene, we can compare the high spatial frequency content of regions in each image and determine the correct focus position. Recovering depth in this manner is inherently a local operation, and can be done efficiently using a pipelined image processor. Laplacian and Gaussian pyramids are used to calculate sharpness maps which are collected and compared to find the focus position that maximizes high spatial frequencies for each region.

1. Introduction

A fundamental foci of study in computer vision is the task of building a three dimensional object-oriented description of a scene given only two dimensional camera images. The availability of absolute depth cues has the potential to greatly reduce the ambiguity present in many image interpretation algorithms. Traditional methods of depth recovery include sonar and laser range finders, which are limited in dynamic range, and image processing methods such as motion parallax and stereo vision, which are limited in their ability to provide low-level depth cues due to correspondence ambiguity.

Depth cues can, however, be recovered by non-intrusive methods without any correspondence problems through the use of optical rangefinding techniques. In particular, depth can be computed from focus, by analyzing a scene across varying focus positions and finding the correct focus for each object in the scene. For a given camera lens there is a one-to-one mapping between focus position and the range of the point in correct focus, thus correct focus position determines depth. We have developed a method of depth recovery that yields unambiguous and complete depth information in near real time using a pyramid-based analysis of focus.

2. Sharpness Criterion

Central to our algorithm is quantifying a notion of proper focus. Intuitively, proper focus is equated with sharpness, or the presence of high spatial frequencies. Conversely, "smoothed", "blurry" and "out of focus" are synonymous. We assume a model of a de-focused image as the convolution of the properly focused image with a low-pass spatial frequency filter, specifically a two dimensional Gaussian [1].

A defocused image D is thus the convolution of a properly focused image I of the same scene with a Gaussian whose width w increases as the focus position of D moves away from that of I.

$$D = I \circ G_w$$

In the frequency domain this is equivalent to multiplying the power spectrum with a Gaussian transfer function (figure 1.)

$$P(D) = P(I \circ G_w) = P(I) \cdot G_w$$

Thus a de-focused image will have a lower high-frequency response than a properly focused image of the same scene. If we examine a particular high frequency interval of the power spectrum across changing focus position, it will reach a maxima at the position of correct focus (figure 2.) We can use a band

CH2605-4/88/0000/0504$01.00 © 1988 IEEE

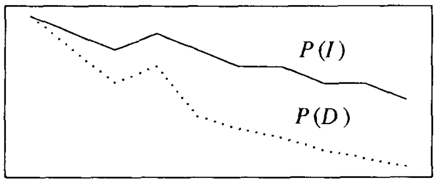

Figure 1. Power Spectrum of Focused and De-Focused Region.

pass filter in the frequency domain to compare the relative sharpness of a scene across different focus positions.

Some work has been done into inferring depth information from a single partially focused image [3]. In this work a sharpness criterion is applied to a collection of image features (edges, lines) and relative depth is calculated based on their sharpness. Blurry features are assumed to be farther from the plane of proper focus than sharp features. This method yields two solutions, since the relative depth can either lie beyond the plane of proper focus or between it and the camera lens. A more critical flaw is that this method assumes all image primitives are perfectly sharp, and that all degradation in the output of the sharpness criterion is only due to de-focusing effects. This method will be confused by scenes that naturally contain edges and lines in varying degrees of sharpness.

The response of a sharpness criterion on a single image is insufficient to determine unambiguous depth information. Given the Gaussian model of lens defocusing, however, we know that the sharpness criterion will be greater for a properly focused region than for a poorly focused image of the same region, irrespective of the actual content of the region. By examining several images of the same scene taken at various focus positions we can find the focus position that maximizes the sharpness criterion for a region, and conclude that focus position to be "proper focus" for the region.

In cases where there is no single peak, or in which the sharpness criterion fails to exceed a minimum threshold, we return no depth information. This method yields depth information wherever there is sufficient high spatial frequency textures and edges in the image. Since we are comparing across the entire range of focus positions, there is no hill climbing problem of false peaks or local maxima.

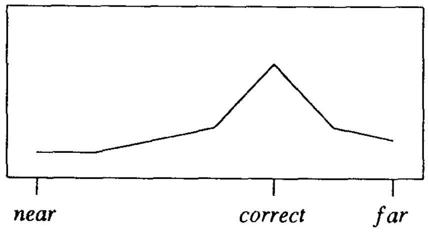

Figure 2. High frequency response of a region vs. focus position

3. Sharpness Maps

Our discussion so far has been under the explicit assumption that the scene is already partitioned into regions with single objects, which presupposes the ability to segment those objects. But the goal of this process is to provide depth cues for segmentation, so we cannot reasonably assume any a priori knowledge about the objects, including their location. Instead we arbitrarily partition the scene into an array of square regions, and then attempt to minimize the error due to overlapping image features when computing proper focus for each region. To compute the sharpness criterion for each region, we convolve the image with a band pass filter, and sum the filter response over each region. The resulting array of sharpness criterion values is termed a "sharpness map."

The processing architecture for calculation of Gaussian and Laplacian pyramids developed at the SRI David Sarnoff Research Center has proven to be well suited to the task of computing sharpness maps. Using a computationally efficient algorithm of smoothing and sampling, Gaussian and Laplacian pyramids of a source image can be constructed in real time using pipelined image processing hardware. A Gaussian pyramid represents a source image I in decreasing levels of size and spatial resolution, effectively applying a low pass spatial frequency filter of increasing width and shrinking the result to compute each successive level of the pyramid

$$G_0(I), G_1(I), \cdots, G_n(I)$$

$G_0(I)$ is identical to I in content and size; each higher level of the pyramid reduces the height and width of the image by half. A Laplacian pyramid

$$L_0(I), L_1(I), \cdots, L_n(I)$$

can be constructed as the difference in successive levels of the Gaussian pyramid, and thus represents I bandpassed with filters of increasing spatial fre-

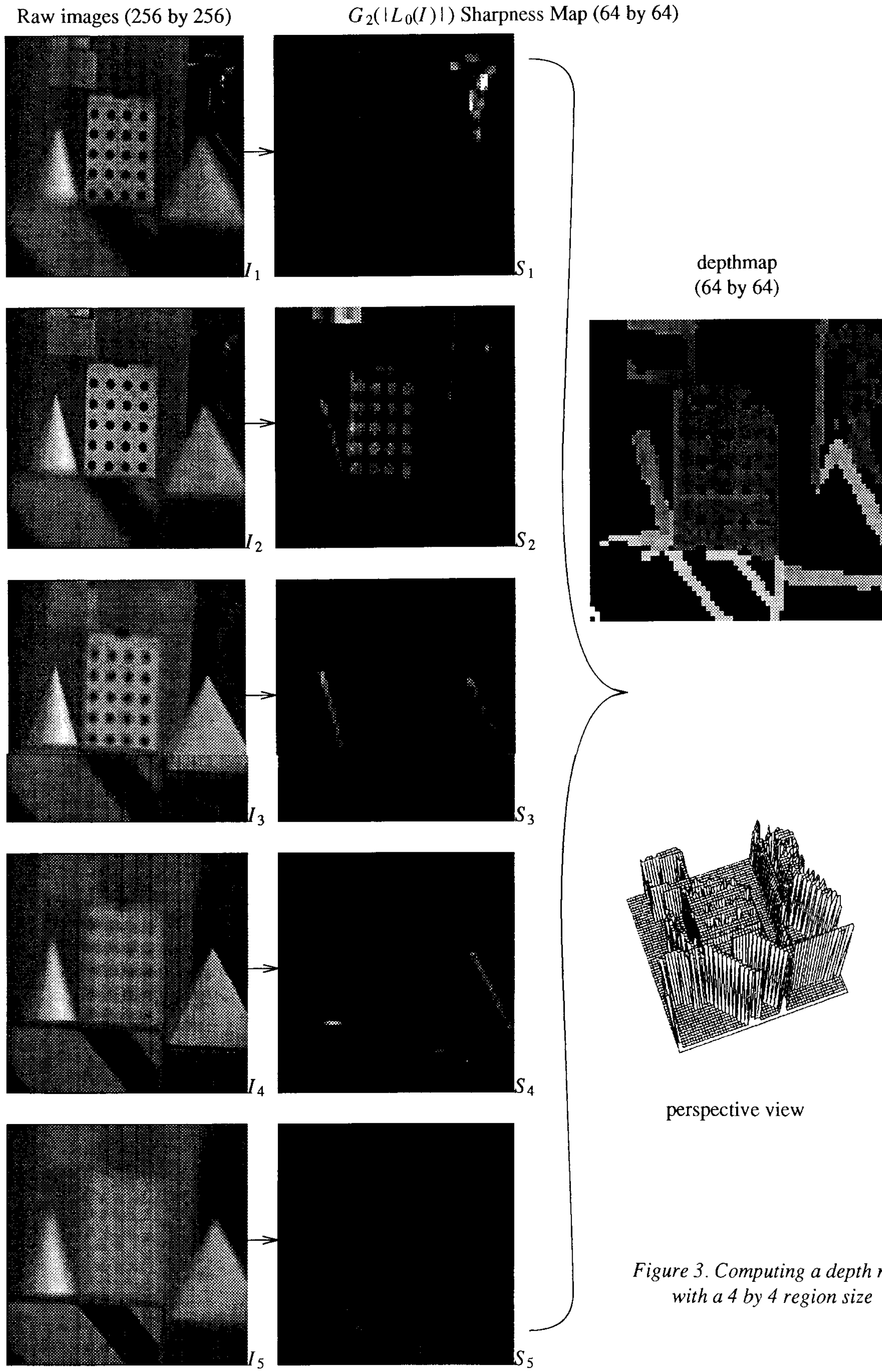

Figure 3. Computing a depth map with a 4 by 4 region size

quency. Burt and Adelson have shown that the pyramid

$$L_0(I), L_1(I), \cdots, L_{n-1}(I), G_n(I)$$

is sufficient to reconstruct I, and is useful as a compaction method since it can be represented in less memory. [4]

From the standpoint of calculating our sharpness map, recall that our sharpness criterion is an integral over a band of the power spectrum of the source image, which we then average over each region to compute an sharpness value on the map. Since the absolute value of the Laplacian represents the integral of one octave in the power spectrum, we can to use $|L_k(I)|$ as the first step towards computing our sharpness map. If the region size is larger than the area a pixel of L_k covers on the original image, we average all values of $|L_k(I)|$ that cover the region. Since we are using square regions, we can do this efficiently by smoothing and shrinking the $|L_k(I)|$ image with a Gaussian pyramid

$$G_0(|L_k(I)|), G_1(|L_k(I)|), \cdots, G_n(|L_k(I)|)$$

Our sharpness map S is thus defined as a function with parameters i and k of an image I with m by m pixels,

$$S(I) = G_i(|L_k(I)|)$$

yielding a map with $\frac{m}{2^{i+k}}$ by $\frac{m}{2^{i+k}}$ sharpness criteria values each of which sample the kth spatial frequency octave of a 2^{i+k} square region of I.

4. Recovering a depth map

Given a set of sharpness maps of images taken at various focus positions it becomes relatively straightforward to compute a depth map. Taking the sharpness value for a single region from each sharpness map, we perform the comparison algorithm discussed above, either interpolating or directly finding a maxima, yielding D, the map of "proper" focus position for each region (figure 3):

$$D(I_1, I_2, \cdots, I_n) = M(S(I_1), S(I_2), \cdots, S(I_n))$$

where $M(I_1, \cdots, I_n)$ returns for each pixel the index position $1 \cdots n$ of the image with the largest value at that pixel. This map of proper focus positions can be returned as a relative depth map, or translated to absolute distance according to lens specifications or prior calibration.

5. Sharpness Map Parameters

There are, of course, trade-offs in the choice of the parameters that determine the sharpness map. The depth resolution of this method is determined by the number of different focus positions sampled, limited by the depth of field of the lens and the sensitivity of the sharpness criteria.

The sensitivity of the sharpness criteria can roughly be equated to the width of the peak found when examining the criteria response across images of a sharp edge taken at different focus positions. It will depend on what range of the power spectrum the sharpness criterion samples; as k increases, the range of frequencies sampled by $G_i(|L_k(I)|)$ decreases, and thus becomes less sensitive to the low-pass filtering caused by defocusing. $G_i(|L_0(I)|)$ will provide the highest possible depth resolution, but also requires the sampling the largest number of different focus positions to find the peak.

The horizontal and vertical resolution of the resulting depth map is determined by the number of regions in the sharpness map; computing a depth map based on regions 8 pixels square for an original image that is 256 by 256 in size would result in a depth map with 32 by 32 regions. By the nature of pyramid representations, and frequency filtering in general, each pixel in $|L_k(I)|$ reflects the response of the power spectrum over a 2^k pixel square region of I. Thus, the smallest possible region a sharpness map based on $|L_4(I)|$ could use would be 16 pixels square; using $|L_0(I)|$, a sharpness map could theoretically have regions equal to the pixel size of I.

As discussed earlier, an underlying assumption of our work is that the sum of $|L_k(I)|$ is greater for a sharp (focused) edge than for a smooth (de-focused) edge. However, it is not the case that it will be *everywhere* greater; if the region size is small enough to sub-sample the Laplacian response on an edge, the sharpness criteria may be erroneous. Ideally, we need to threshold sharpness values such that the entire response on a focused edge does lie above the response of a fully de-focused edge. In the above example if we had used a region size of 1 and no threshold on the sharpness criterion, we would be in trouble at the points where $|L_0(I)|$ of the smooth edge is greater than $|L_0(I)|$ of the sharp edge, returning the focus position of the de-focused edge as the proper focus position. Practically, a moderate threshold and region size 3 or 4 times the underlying Laplacian pixel size are usually sufficient to keep these errors from becoming significant.

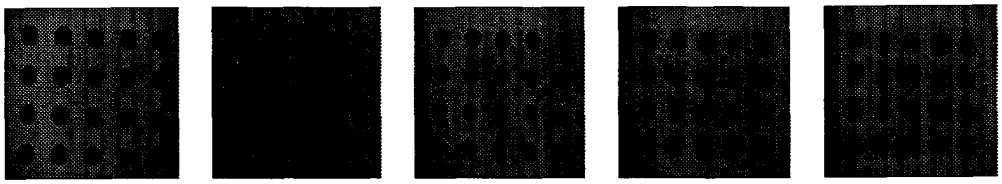

Figure 4. Calibration image sequence.

In the Burt pyramid processing architecture it is equally easy to compute the single image $G_i(|L_k(I)|)$ as it is to compute the pyramid

$$G_1(|L_k(I)|), G_2(|L_k(I)|), \cdots, G_i(|L_k(I)|)$$

Thus it is quite plausible to punt on specifying a value for the sharpness map parameter i, and treat S instead as a sharpness pyramid. In this case the result of sharpness comparison would be a multi-resolution depth pyramid.

6. Compensating for Zoom distortion

In comparing $S(I_f)$ across different focus positions we must account for the zoom effect experienced with changing focus. Since we have cameras with servo-controlled zoom position, we can compensate for this optically, by changing the zoom position with the focus position. Alternatively, we can compensate by calculating a distortion map that yields equivalent image positions for each focus position and applying it to the image before the sharpness map is calculated. Increasing the region size also helps to reduce the zoom effect.

To compute a distortion map for the zoom effect we take images of a test pattern containing an array of circles (figure 4,) and track the center of these circles as they move with focus position. For each circle we fit the observed points into a vector v (figure 5.)

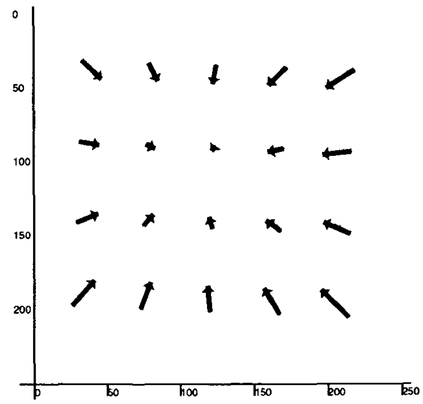

Figure 5. Object movement with changing focus.

We take all the nearly orthogonal intersections of the lines containing these vectors, and compute the most likely lens center to be at the center of these intersection points (figure 6.) Having thus obtained the lens center, we fit the vector lengths as a function of radius, which allows us to predict the pixel change due to changing focus for any point in the scene.

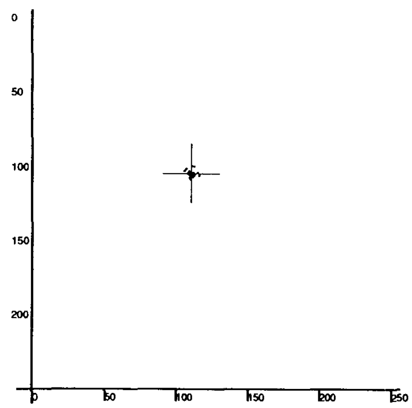

Figure 6. Calculated lens center.

7. Real Time Performance

Using hardware that performs pyramid decomposition, the calculation of $G_n |L_0|$ sharpness pyramids can be done in real time (1/8 second per image). Calculation of maxima can also be done at frame speeds in pipelined image processing hardware. The performance of our current implementation is bounded by the time it takes the focus and zoom motors to move through all focus positions that are to be sampled. This is on the order of one second per focus position, thus we can compute a depth map with 10 levels of depth resolution in roughly 10 seconds.

8. Conclusion

We have presented a simple and efficient method of recovering depth information by analyzing sharpness across different focus positions. There is no correspondence problem as found in stereo vision nor intrusive emissions as from laser rangefinders. Our method yields depth information anywhere there is high frequency information–edges, lines, textures, or any other "interesting" 2D image features. This work shows there is valuable depth information not normally used by static image understanding algorithms that is recoverable using straightforward and unambiguous active perception techniques.

9. Acknowledgements

We would like to thank the SRI David Sarnoff Research Center for providing the hardware used in this research, and to P. Burt for suggesting the Laplacian as a sharpness criterion. Support for this work was provided under DARPA contract N0014-85-K-0807.

10. References

[1] A. Pentland, "A New Sense for Depth of Field," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-9, pp. 523-531, July 1987.

[2] A. McIvor, "Stochastically Based Vision Algorithms," Ph.D. Thesis, Div. of Applied Science, Harvard Univ., May 1985.

[3] P. Grossmann, "Depth from Focus," *Pattern Recognition Letters,* vol. 5, pp. 63-69, Jan. 1987

[4] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. Communication,* vol. COM-31, pp. 532-540, 1983.