View Selection Strategies for Multi-View, Wide-Baseline Stereo

Hany Farid
University of Pennsylvania

Sang Wook Lee
University of Pennsylvania

Ruzena Bajcsy
University of Pennsylvania

View Selection Strategies for Multi-View, Wide-Baseline Stereo

MS-CIS-94-18
GRASP LAB 373

Hany Farid
Sang Wook Lee
Ruzena Bajcsy

University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

May 1994

View Selection Strategies for Multi-View, Wide-Baseline Stereo

Hany Farid, Sang Wook Lee, Ruzena Bajcsy
{farid, swlee, bajcsy}@grip.cis.upenn.edu
GRASP Laboratory
Department of Computer Science
University of Pennsylvania
Philadelphia, PA 19104-6228

Abstract

Recovering 3D depth information from two or more 2D intensity images is a long standing problem in the computer vision community. This paper presents a multi-baseline, coarse-to-fine stereo algorithm which utilizes any number of images (more than 2) and multiple image scales to recover 3D depth information. Several "view-selection strategies" are introduced for combining information across the multi-baseline and across scale space. The control strategies allow us to exploit, maximally, the benefits of large and small baselines and mask sizes while minimizing errors. Results of recovering 3D depth information from a human head are presented. The resulting depth maps are of good accuracy with a depth resolution of approximately 5mm.

1 Introduction

The major steps in recovering depth information from a pair or sequence of images are: (1) preprocessing, (2) matching, and (3) recovering depth (see [1] for a review of stereo algorithms). The preprocessing stage generally consists of a rectification step which accounts for lens distortion and non-parallel axis camera geometry ([5], [7]). This stage may also consist of an intensity normalization step. The process of matching is the most important and difficult stage in most stereo algorithms. The matching process determines correspondence between "features" that are projections of the same physical entity in each view. Matching strategies may be categorized by the primitives used for matching (e.g. features or intensity) and the imaging geometry (e.g. parallel or non-parallel optical axis). Once the correspondence between "features" has been established, calculating the depth is usually a straight forward computation dependent on the camera configuration and optics.

One of the most common stereo reconstruction paradigms is matching image features from two parallel axis views (see [6] for a review). This method provides a disparity value d for matched pairs of points for each point in either the left or right image. The depth z can then

be recovered by the well known equation: $z = \frac{fb}{d}$, where, f is the focal length of the pin-hole camera model, and the baseline b is the distance between the two focal points of the cameras. This approach to recovering stereo is attractive because of its simplicity; however, recovering an accurate, dense 3D depth map with this procedure has proven to be a formidable task.

The desired properties of a 3D depth map may vary depending on the application. For example, mobile robots performing obstacle avoidance are likely to be more interested in a fast, coarse estimate of depth, whereas a robot interested in manipulating its environment is more likely to be interested in generating a more accurate depth map. The interests of the authors are in the area of telepresence [2]. Telepresence may described as a system where participants wear a head-mounted display to look around a remote environment. The surface geometries of the remote environment are continuously sensed by a multitude of cameras mounted along the ceiling and walls, from which depth maps are extracted. For the application of telepresence the desired properties of a 3D depth map are (1) high density, (2) high resolution, (3) good localization; of features and (4) minimal errors due to specularity, occlusion, and camera calibration.

Two standard parameters that most stereo algorithms vary are the *baseline*, and the *mask size* over which correlation is performed. Varying these parameters effects different properties of the recovered depth map (See Figure 1). In particular, a large mask size will result in a high density depth map (i.e. good recovery in the absence of "features") but poor localization of features in the 'x' and 'y' dimension. A large baseline line allows for high resolution in the 'z' dimension, but increases the likelihood of errors due to occluding boundaries and repetitive patterns in the scene.

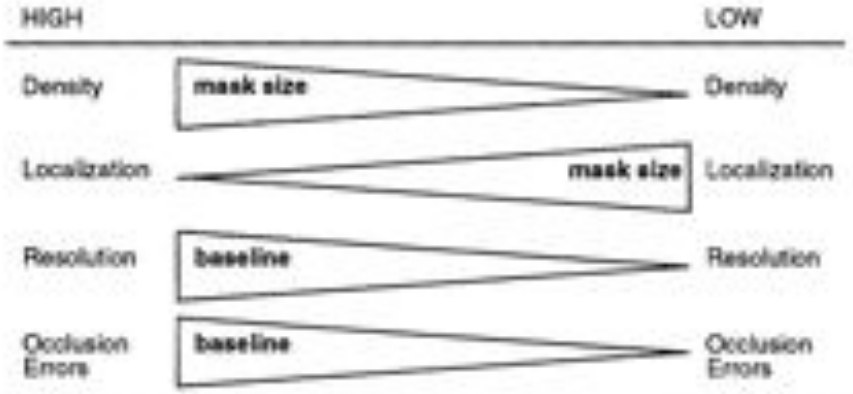

Figure 1: Tradeoffs between the baseline (distance between successive views) and the mask size (size of sampling neighborhood in which feature matching is performed). For example, a large mask size results in high density but low localization.

In this paper, a stereo algorithm is presented which attempts to exploit, maximally, the benefits of small and large baselines and mask sizes. In particular, a *multi-baseline, coarse-to-fine* approach to stereo is adopted; where several closely spaced views are taken (multi-baseline) and matching across these views is done for several different mask sizes (coarse-to-fine). The use of several views and mask sizes introduces a need for more sophisticated *matching* and *combination* strategies. Several such control strategies are introduced for matching across the mutli-baseline which greatly reduce errors due to repetitive patterns and false matches that arise from occluding boundaries. Control strategies are also introduced for combining information

across varying mask sizes which lead to dense, high resolution depth maps.

The algorithm presented in this paper is based on the original multi-baseline algorithm in [4, 3]. A description and discussion of this algorithm is presented in Section 2. Section 3 introduces a stereo algorithm along with the control strategies for combining views across the mutli-baseline and scale space. Section 4 presents experimental results of recovered depth maps from a human subject along with a brief analysis of the effectiveness of the control strategies used in the algorithm. A brief discussion and future research directions is given in Section 5.

2 Multi-Baseline Stereo

Most stereo algorithms perform feature matching across only two views. This basic paradigm is vulnerable to false matches in the presence of repetitive patterns (see Figure 2).

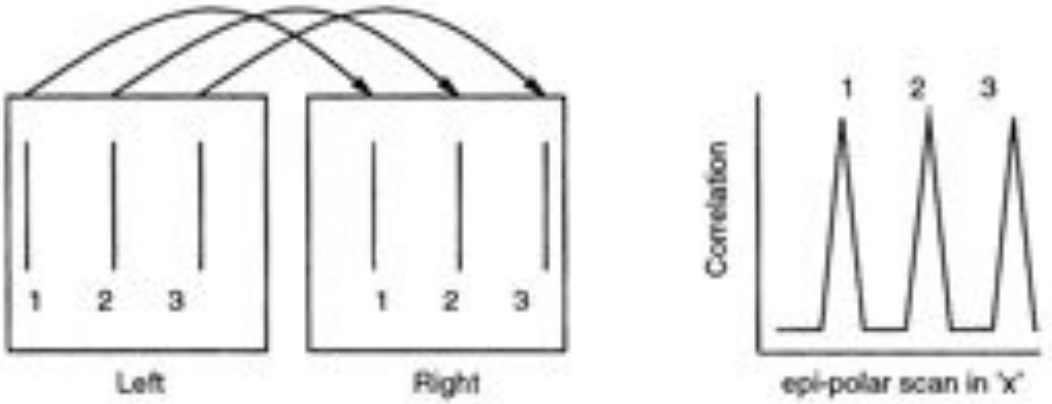

Figure 2: Two images with repetitive patterns (vertical lines). Matching performed from the left to the right image for a point on line 1 results in a false match (i.e. two spikes appear in the correlation curve).

The original multi-baseline stereo algorithm described in [4, 3] was introduced in order to avoid false matches due to repetitive patterns. The basic idea is to compute correlation curves between a reference image (the left or right-most image) and a series of equally space images with successively larger baselines. The correlation curves are then added and the correct match is taken to be at the global minimum of this curve. This mutli-baseline eliminates false matches due to repetitive patterns having a frequency greater than the shortest baseline. This method, however, is limited in that the *largest* baseline (i.e. distance between left and right-most cameras) must remain relatively small to avoid the introduction of errors from occluding boundaries. The consequence of a small baseline is low depth resolution in the recovered depth maps.

In this paper, we introduce *view-selection control strategies* which allows us to take advantage of the multi-baseline approach for *large* baselines, thus allowing us to recover high resolution depth maps. Several other control strategies are also introduced which allow us to combine depth maps computed from varying mask sizes, allowing us to recover accurate and dense depth maps.

3 View-Selection Strategies for Wide-Baseline Stereo

A *multi-baseline, coarse-to-fine* intensity based algorithm for recovering 3D depth information from a sequence of images is described in this section. This approach utilizes several control strategies which allows us to exploit, maximally, the benefits of both small and large baselines and mask sizes.

3.1 Intensity Matching

In order to recover dense depth maps, intensity matching, as opposed to feature matching, is used in the stereo algorithm presented in this section. In particular, matching correlation error is given as the sum of absolute value of differences of intensities over a sampling window:

$$\sum_{j=1}^{n} \frac{| I_j - \hat{I}_j |}{n} \tag{1}$$

where, I_j and $\hat{I}_j$ are the intensity values in the images being matched and n is the the number of pixels over which correlation is performed (i.e. mask size). This correlation measure was chosen over the frequently used *sum of squared differences* for computational considerations.

3.2 Wide-Baseline Stereo

Most feature based matching algorithms generally rely on only two images to recover a 3D depth map [6]. This method is limited in that a single fixed baseline must be chosen; thus either good resolution or minimizing errors due to occlusion may be sacrificed. A fixed, large baseline makes matching more difficult due to an increased chance of false matches introduced by occluding boundaries. In addition, a large baseline is especially problematic in the presence of repetitive patterns [4, 3] (e.g. vertical stripes on a shirt). Selection of a small baseline reduces the chance of false matches but at a cost of poorer depth resolution.

In order to take advantage of the benefits of using a small and large baseline, matching may be performed over a sequence of images [4, 3]. For example, Figure 3 depicts a seven image sequence where the distance between successive frames is small while the "full baseline", the distance between the left and right most images (L_3 and R_3), is large. There are several strategies that may be adopted for matching across such a sequence of images; below, we present one such approach.

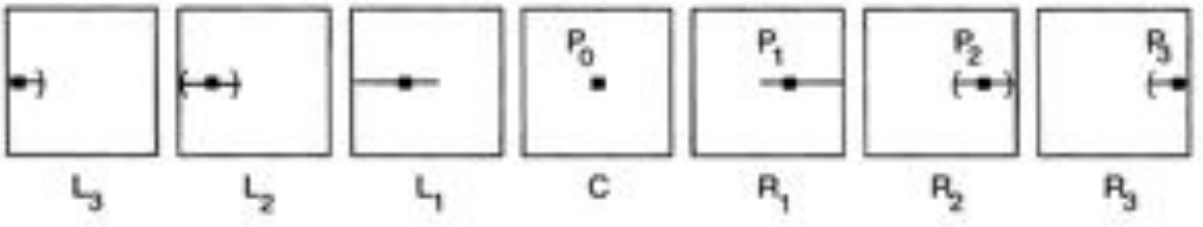

Figure 3: Matching across an image sequence.

Whereas the original multi-baseline stereo algorithms [4, 3] performs correlation to the left or right most image in a sequence of images, the algorithm described here correlates to the center image in the sequence. Correlating to the center view, in effect, reduces the baseline by a factor of two thus making errors due to occlusion, etc. less likely. The benefit of the full baseline is, later, partially recovered as will be described below.

Consider for the moment only the right half of the image sequence in Figure 3 (images C through R_3). The matching point of a point P_0 in image C can be determined in image R_1 by searching along an epi-polar line. [1] Let the point P_1 be the matching point in image R_1. The matching point for P_1 in image R_2 can then be determined by searching about a epi-polar line centered at the projection of P_1 in image R_2. Finally, the matching point for P_2 in image R_3 can be determined by searching about a epi-polar line centered at the projection of P_2 in image R_3. The disparity for P_0 is then simply $P_2^x - P_0^x$, where P_i^x is the x component of the point P_i. [2] In order to avoid errors due to occlusion, if the correlation error of a point in image P_i is above a pre-defined threshold, then the previously matched point P_{i-1} is directly projected into the last image in the sequence.

The projection of points is trivial given a known distance between neighboring images in the sequence. Given an image sequence with n images, a point P_i in image i is projected into image $i+1$ as follows:

$$P_{i+1} = P_i * ((i+1) * \frac{n}{2})/(i-n) \tag{2}$$

Errors in the projection can be compensated for by increasing the search neighborhood about the projection point.

The process of computing disparity for a single point is repeated for each point in image C, resulting in a disparity map relating points in image C to those in image R_3. The process is then repeated to compute a disparity map relating points in image C to those in image L_3.

In order to take advantage of the full baseline (image L_3 to R_3), it is necessary to "combine" the left and right disparity maps. In an ideal world these maps would be identical and simply adding them would suffice. However, due to occlusions, noise, intensity variations, false matches, etc. this approach is unrealistic and results in a large number of errors. As such, a simple "combination rule" to combine the left and right disparity map is adopted:

$$\begin{gathered}\text{if } (|D_L - D_R| \ < \ \varepsilon_D \text{ and } |C_L - C_R| \ < \ \varepsilon_C) \text{ then } D_F = (D_L + D_R) \\ \text{else if } (C_L < C_R \ ) \text{ then } D_F = 2 * D_L \\ \text{else } D_F = 2 * D_R\end{gathered}$$

where, D_L and D_R corresponds to the left and and right disparity maps, respectively, C_L and C_R correspond to the left and right correlation errors, respectively and D_F corresponds to the final disparity value. ε_D and ε_C are pre-defined thresholds set to a value of 1 in the results presented in section 4. These two thresholds dictate the error tolerance between the left and right disparity maps.

[1] The matching point is determined by correlating intensity values over a fixed size mask and selecting the point with a minimum correlation. Correlation is computed as a sum of absolute value of differences.

[2] This assumes parallel axis camera geometry.

To this point correlation has been performed only over a single mask size, the following section describes how varying mask sizes are incorporated into this algorithm.

3.3 Coarse-to-Fine

In order to benefit from the properties of correlating over a large and small mask size, disparity maps are computed for a number of mask sizes. In particular, using the process described in the previous section, disparity maps are computed for mask sizes ranging from 3×3 to 15×15. Associated with each of these disparity maps is a correlation map, which associates a correlation value with each point in the image. The final disparity map is computed by initially setting all disparity values to be that of the coarsest map (15×15 mask). Each point in the final disparity map is then updated through the smaller mask sizes as long as the correlation of a smaller mask is less than or equal to the correlation of a larger mask.

The algorithm described in the previous two sections is outlined, in pseudo-code, in Appendix A.

4 Results

4.1 Experimental Setup

In order to obtain a sequence of images while insuring parallel axis motion, a CCD camera (Sony XC-77RR, 25mm lens) is mounted on the end effector of a PUMA 560. An image sequence is then obtained by repeatedly moving the PUMA a fixed distance horizontally in front of an object, and digitizing an image at each step (Figure 4).

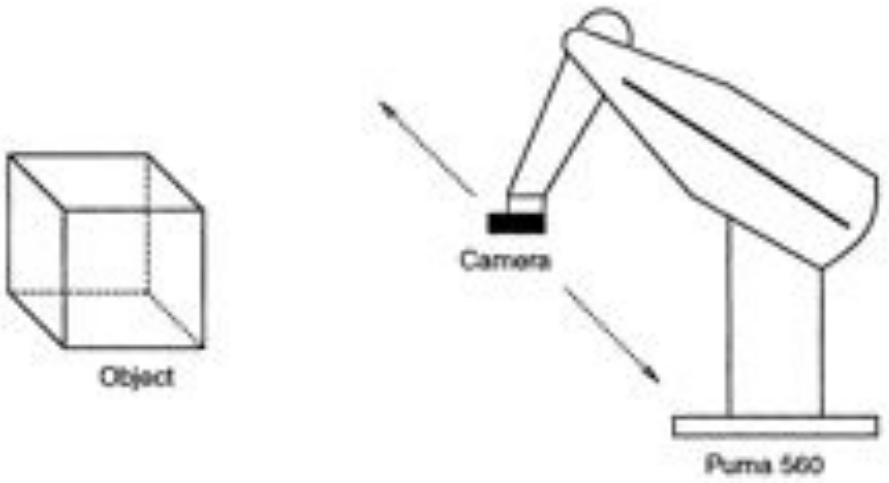

Figure 4: Experimental setup.

4.2 Depth Maps

A five image sequence of the upper torso of a human subject was taken. The camera was translated 7cm between successive views, giving a full baseline of 28cm. The subject was approximately 1m from the camera. The stereo reconstruction algorithm described in the previous section was run on the subsampled 256 × 256 images (images were originally 512 × 512).

Figure 5: Image sequence.

Figures 6, 7 and 8 show the resulting rendered depth map. Figure 6 shows the rendered depth map (clockwise) rotated by 45° about the x-axis, about the y-axis, rotated by −45° about the x-axis and about the y-axis. Figure 7 shows the same views where the depth map was upsampled to 512 × 512 using bicubic spline interpolation. Figure 8 shows the upsampled views rotated by ±90° about the y-axis. The upsampled depth map has a resolution of approximately 5mm.

4.3 Analysis

It is not clear what the "best" method is for combining left and right disparity maps and for combining disparity maps across scale space. The method presented in this paper is based on some simple heuristics; further study into this area is certainly warranted. In order to aid in this study we have performed some preliminary analysis on how effective the various control strategies introduced in this paper are.

Table 1 shows the fraction of time the disparity map for a given mask size was obtained from the left, right or both disparity maps (see Table 1). Figure 9 shows, for the 15 × 15 mask size, what region of the image used the left, right, and both disparity maps; in this figure, dark gray corresponds to the left disparity map only, gray corresponds to a combination of both the left and right disparity maps, and white corresponds to the right disparity map only. Note it is at the occluding boundaries (i.e. left and right sides of the head) that the disparity is computed from only the right and left disparity map, respectively. From this figure it is clear why the control strategy introduced in this paper helps to avoid errors due to occluding boundaries.

Table 2 shows the fraction of the time in which each mask size was used in generating the final disparity maps. Recall that the final disparity map is initially set to the coarsest disparity map (i.e. 15 × 15) and is updated through different mask sizes as long as the correlation from a smaller mask size is less than that of a larger mask size. We have found that, as expected, smaller mask sizes are more effective in regions where "features" are present and in regions of high texture, while larger mask sizes are more effective in more uniform regions. Combining

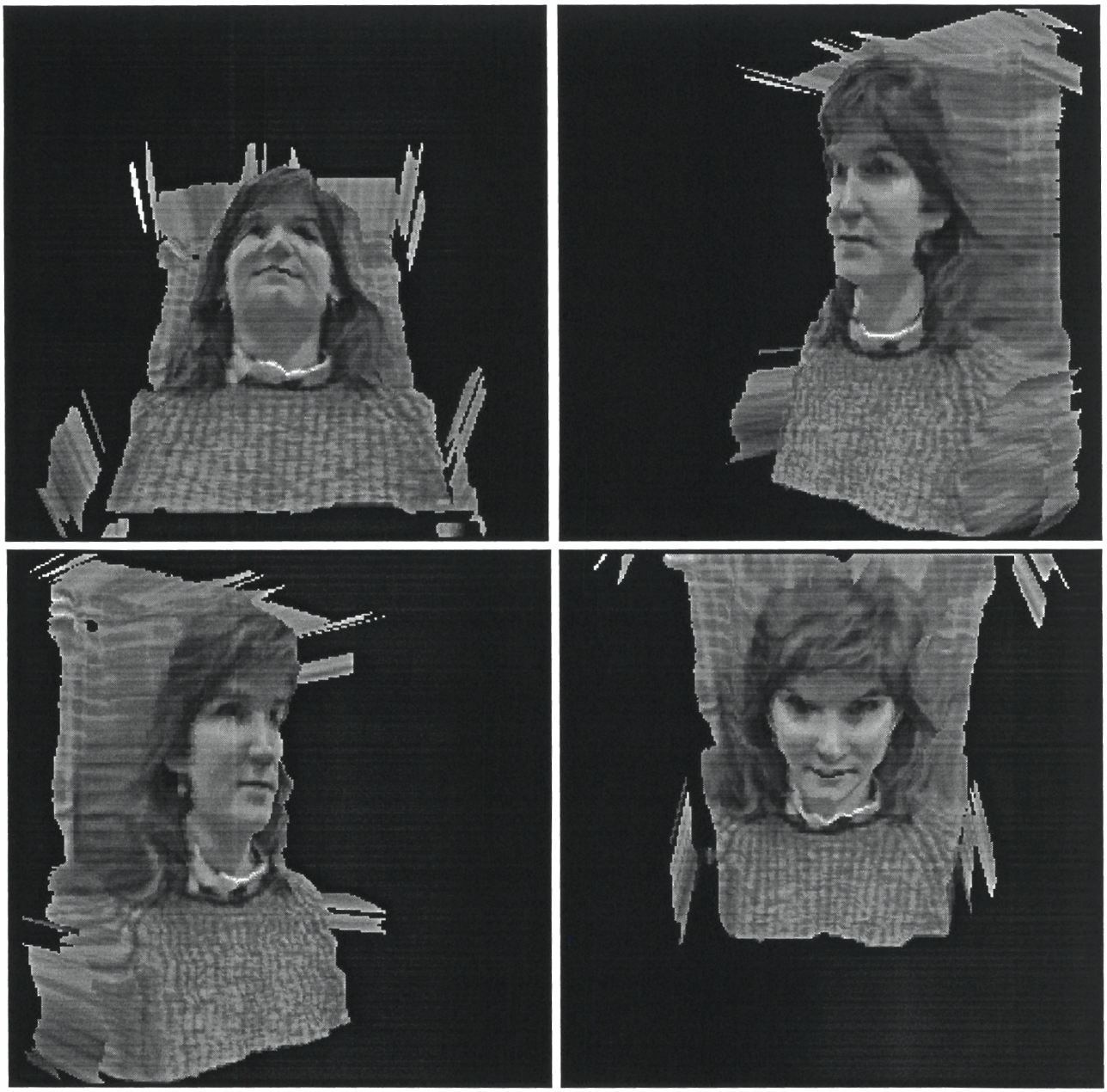

Figure 6: Rendered depth map.

Figure 7: Rendered depth map upsampled to 512×512 using bicubic interpolation.

Figure 8: Rendered depth map upsampled to 512 × 512 using bicubic interpolation.

Table 1: Combining left and right disparity maps.

Mask Size	% Using Left Disparity Map	% Using Right Disparity Map	% Using Both Disparity Maps
3 × 3	23.7	26.1	44.5
5 × 5	20.5	22.0	53.1
7 × 7	19.7	20.9	55.8
9 × 9	19.5	20.3	57.5
11 × 11	18.9	20.1	59.1
13 × 13	18.5	19.7	60.5
15 × 15	18.3	19.7	61.6

Figure 9: Combining left and right disparity maps. Dark gray corresponds to the left disparity map only, gray corresponds to a combination of both the left and right disparity maps, and white corresponds to the right disparity map only.

disparity maps across different masks sizes results in dense 3D depth maps with few errors due to false matches.

A more thorough investigation into the "control strategies" is currently under investigation. We, however, note that the simple strategies described above generates depth maps with fewer errors due to occluding boundaries and false matches than the original multi-baseline algorithm described in [4, 3]. Figure 10 shows the results of running our implementation of the original multi-baseline algorithm on the image sequence shown in Figure 5.

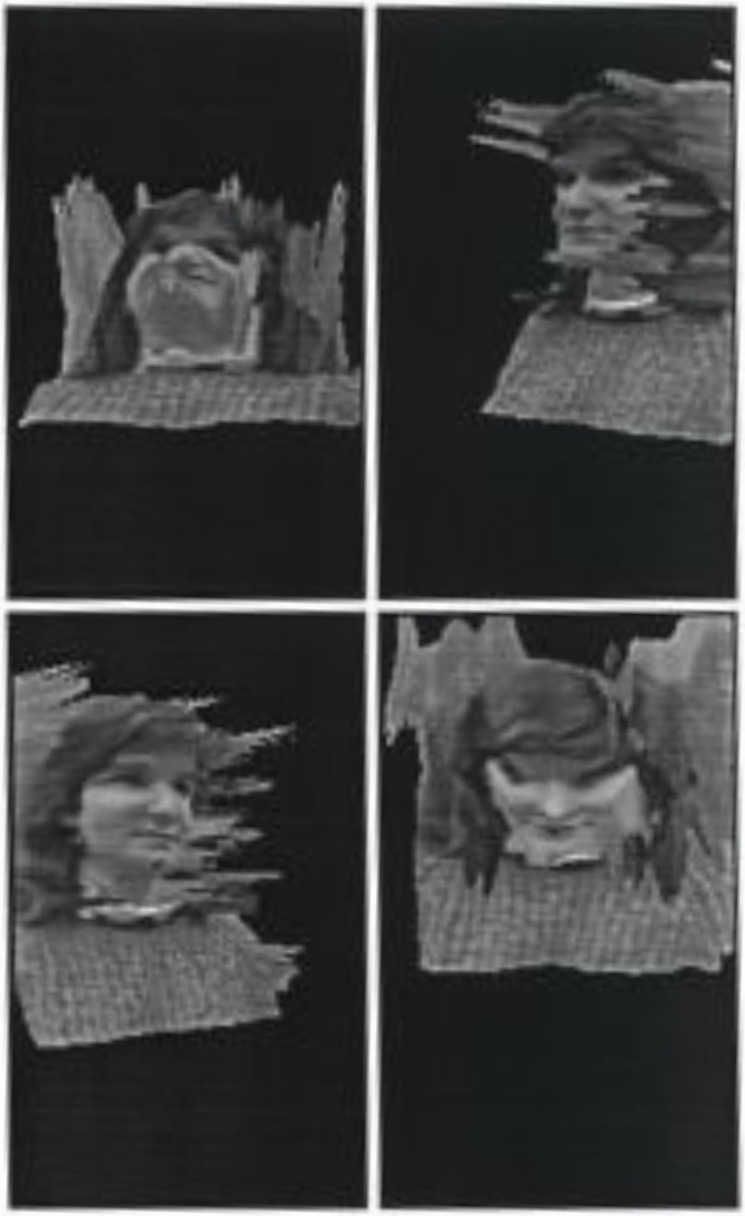

Figure 10: Depth map from our implementation of Kanade's multi-baseline algorithm. Correlation was performed on the left-most image in a 5 image sequence using a 9×9 mask.

Table 2: Combining disparity maps across varying mask sizes.

Mask Size	Percent
3×3	17.96
5×5	10.33
7×7	7.14
9×9	5.50
11×11	4.59
13×13	4.08
15×15	50.41

The original multi-baseline stereo algorithm correlates to the left-most image in the sequence. Correlation curves are computed between the left-most image and each subsequent image in the sequence. The final disparity is gotten from the minimum of the sum of all the correlation curves. The algorithm was run with a fixed mask size of 9×9. Since a coarse-to-fine approach was not incorporated into this algorithm we do not expect as good resolution. Note, however, that the errors due to occlusion (right side of the head) are much more severe in this image than in Figures 6, 7, and 8. The errors in the chin are most likely due to false matches and was observed in our stereo algorithm before the various control strategies were introduced.

5 Discussion

This paper introduces a multi-baseline, coarse-to-fine stereo algorithm with several control strategies for combining disparity information across the image sequence and combining depth information across scale space. These control strategies allow us to benefit, maximally, from the large and small baseline and mask sizes while introducing few errors. The resulting depth maps are of high density, high resolution (approximately 5mm) and low errors due to occlusion, repetitive patterns, and false matches.

Future directions currently include expanding the mutli-baseline to incorporate vertical as well as horizontal camera displacements (i.e. a "grid" of cameras). In addition we are interested in investigating methods to integrate depth maps taken from different perspectives. The addition of more cameras and views will increase the importance of the type of "control" issues introduced in this paper. Our future work includes further investigation into this aspect of stereo reconstruction.

We are also currently beginning a thorough analysis of the errors introduced at each stage of the stereo reconstruction process. In particular, camera calibration, image rectification (due to lens distortion and non-parallel axis camera geometry), and the matching stage of the stereo algorithm. This type of analysis will hopefully result in a better understanding of where errors are introduced into the depth map and how they may be avoided.

6 Appendix A

Below is the multi-baseline, coarse-to-fine stereo algorithm outlined in Section 3. The algorithm was implemented in 'C' on a Sparc platform.

```
/* compute disparity maps for varying mask sizes */
for m = 3 to 15 {                                        /* mask size = m x m */

     /* compute right disparity map */
     for each pixel (x,y) in center image
          for I = 1 to n {        /* images in right half of image sequence */

               find matching point (x1,y1) in image I          /* correlate */
               c = (min correlation)/m^2           /* normalize correlation */
               project (x1,y1) into image I+1
          }
          right_disparity = x1 - x
          right_correlation = c

     /* compute left disparity map */
     for each pixel (x,y) in center image
          for I = 1 to n {         /* images in left half of image sequence */
               find matching point (x1,y1) in image I          /* correlate */
               c = (min correlation)/m^2           /* normalize correlation */
               project (x1,y1) into image I+1
          }
          left_disparity = x1 - x
          left_correlation = c

     /* combine left and right disparity map */
     if( | right_correlation - left_correlation | < Ec &&
         | right_disparity - left_disparity | < Ed ) {
          disparity_m = left_disparity + right_disparity
          correlation_m = (left_correlation + right_correlation)/2.0
     } else if( left_correlation < right_correlation ) {
          disparity_m = 2 * left_disparity
          correlation_m = left_correlation
     } else {
          disparity_m = 2 * right_disparity
          correlation_m = right_correlation
     }
}

/* combine disparity maps across varying mask sizes */
final_disparity = correlation_15     /* set final disparity to coarsest map */
final_correlation = correlation_15

for m = 13 to 3 {
     if( correlation_m < final_correlation ) {
          final_disparity = correlation_m
          final_correlation = correlation_m
```

```
    }
}

/* compute depth */
depth = (focal length)*(baseline) / final_disparity

/* post-processing */
smooth (gaussian) final depth map
```

References

[1] U. Dhond and J. Aggarwal. Structure from stereo - a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, 1989.

[2] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, and H. Farid. Virtual space teleconferencing using a sea of cameras. *Unpublished Results*.

[3] T. Kanade. A multiple-baseline stereo.

[4] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[5] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, 1987.

[6] J. Weng, N. Ahuja, and T. Huang. Matching two perspective views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):806–825, 1992.

[7] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):965–980, 1992.