# 3D reconstruction from images taken with a coaxial camera rig

Richard Kirby* and Ross Whitaker
University of Utah School of Computing, 50 S. Central Campus Drive Room 3190
SLC UT 84112

## ABSTRACT

A coaxial camera rig consists of a pair of cameras which acquire images along the same optical axis but at different distances from the scene using different focal length optics. The coaxial geometry permits the acquisition of image pairs through a substantially smaller opening than would be required by a traditional binocular stereo camera rig. This is advantageous in applications where physical space is limited, such as in an endoscope. 3D images acquired through an endoscope are desirable, but the lack of physical space for a traditional stereo baseline is problematic. While image acquisition along a common optical axis has been known for many years; 3D reconstruction from such image pairs has not been possible in the center region due to the very small disparity between corresponding points. This characteristic of coaxial image pairs has been called the unrecoverable point problem. We introduce a novel method to overcome the unrecoverable point problem in coaxial camera rigs, using a variational methods optimization algorithm to map pairs of optical flow fields from different focal length cameras in a coaxial camera rig. Our method uses the ratio of the optical flow fields for 3D reconstruction. This results in accurate image pair alignment and produces accurate dense depth maps. We test our method on synthetic optical flow fields and on real images. We demonstrate our method's accuracy by evaluating against a ground-truth. Accuracy is comparable to a traditional binocular stereo camera rig, but without the traditional stereo baseline and with substantially smaller occlusions.

**Keywords:** Stereo endoscope, stereo borescope, 3D reconstruction, variational methods, coaxial camera rig, depth from zooming

## 1. INTRODUCTION

3D reconstruction from image pairs taken from two different perspectives is one of the most active areas of research in computer vision [6, 13, 14]. The most common two-camera rig is the binocular stereo rig where the cameras are oriented with their optical axes parallel and separated by a baseline. 3D reconstruction from images taken with a binocular stereo rig typically uses pixel pairs (one from each camera), that are a projection of the same point in the scene [5, 16]. When the camera geometry is known, depth can be estimated using the disparity between the pixel pairs. The resolution of the depth estimate is a function of the pixel size, distance to the scene, and the baseline between the cameras. The larger the baseline, the higher the resolution of the depth estimate. However, larger baselines create two well know problems: 1) The larger the baseline, the greater the likelihood and extent of occlusions (areas of the scene where one camera cannot see what the other camera sees) and 2) the larger the baseline, the larger the camera rig's horizontal form factor.

There are computer vision applications where the form factor of a traditional binocular stereo baseline is problematic, most notably in applications requiring that the camera rig be inserted into a small opening, like the barrel of an endoscope or borescope or in applications where the surface being analyzed is so close to the cameras that sufficient overlap between images is impossible [7]. Traditional binocular stereo endoscopes exist, but either the cameras are so close together that the depth resolution is low, or the instrument is too large for some applications.

One alternative to a traditional binocular stereo rig is a coaxial [8] camera rig which is, in essence, simultaneous depth from zooming [12]. In this type of camera rig, images are taken at two different focal lengths along the same optical axis. This creates a disparity which is a function of both the distance to the point in the scene and the distance that the pixel under evaluation is from the optical center of the camera. This type of camera rig produces results similar to a traditional binocular stereo rig near the edges of the images, but in the center region, the disparities are too small to produce acceptable depth resolution.

*email: richard.kirby@kspresearch.com; phone: 1-435-503-2078; www.kspresearch.com

In this paper we introduce a novel automated method for finding depth in image sequences taken with a coaxial camera rig by using the optical flow fields. We apply the technique to both synthetic optical flow fields and real images taken with an RGB-RGB coaxial camera rig. In applications where there is sufficient motion between the camera rig and the scene (moving endoscope or borescope) and where the scene exhibits enough texture to produce optical flow, our method finds correspondences between the flow fields and uses the ratio of the flow fields at these corresponding points to estimate depth. The resulting dense depth maps are used to perform 3D reconstruction of the scene with accuracies in the center region of the images (where the unrecoverable point problem prevents reconstruction using image features or pixel intensities) that are similar to those of a traditional binocular stereo rig.

## 2. RELATED WORK

Depth from images taken at different focal lengths along a common optical axis was first proposed by Ma and Olsen [12]. Lavest et al. [11, 10] provide a proof for inferring 3D data from images taken at multiple focal lengths along a common optical axis and models a revolving object. Asada et al. [1] and Baba et al. [2] present a method for doing 3D reconstruction using blur from zoom. Gao et al. [4] present a distance measurement system for mobile robots using zooming. Most recently, Zhang and Qi [15] describe a method for 3D reconstruction from multi-focal length images using a snake-search algorithm.

Traditionally, researchers were interested in 3D reconstruction from images taken along the same optical axis at different focal lengths because it only requires one camera. However, there are several other advantages. Ma and Olsen alluded to the fact that a depth from zoom camera exhibits substantially smaller occlusions than a binocular stereo camera rig with the same baseline. Additionally, there are applications where a stereo baseline is prohibitive (endoscope or borescope) and where the known correspondence point on the optical axis is an advantage to image registration. Finally, where image registration is the ultimate objective of the application (e.g. alignment of images from two different types of sensors without attempting 3D reconstruction), a coaxial camera produces substantially smaller disparity errors in the center region than a binocular multimodal stereo rig.

The coaxial camera rig [8] is equivalent to simultaneous depth from zooming, but instead of changing the focal length of a single fixed camera, two cameras are arranged such that the cameras form images along the same optical axis. This is done by splitting the optical path with a beam splitter and aligning the two cameras such that their optical centers image the same point in the 3D scene. The coaxial camera rig combined with image correspondences derived from perceived motion overcomes the two main problems of depth from zooming. First, simultaneous images taken at two different focal lengths overcomes the stationary scene constraint of depth from zooming. Second, using the flow field to align image pairs overcomes the unrecoverable point problem in the center region that was described by Ma and Olsen.

The use of optical flow instead of image intensities or features to align image pairs has an additional benefit, it permits the alignment of images from different types of image sensors (e.g. RGB and IR). This is because optical flow fields are generally invariant to the wavelength of light being imaged. We explore this capability using a traditional stereo baseline camera rig using multimodal sensors and report our results in a paper submitted concurrently to this one [9].

## 3. VARIATIONAL MODEL

Referring to Figure 1, let $\bar{x}_f \coloneqq (x_f, y_f)^T$, $\bar{x}_b \coloneqq (x_b, y_b)^T$ represent points in the image domain of the front and back cameras. Let $\bar{h}(\bar{x}) \coloneqq$ the disparity between $\bar{x}_f$ and $\bar{x}_b$ such that $\bar{x}_f$ and $\bar{x}_b - \bar{h}(\bar{x}_f)$ represent the same point $\bar{X}(\bar{x}_f) \coloneqq (X, Y)$ in the scene. Let $f_f, f_b \coloneqq$ the focal lengths for the front camera and back cameras and $Z(\bar{x}_f) \coloneqq$ the distance between the optical center of the front camera and a point in the scene corresponding to $\bar{x}_f$, the distance being measured along the optical axis. $b \coloneqq$ the distance between the optical center of the two cameras. $\bar{w}_f, \bar{w}_b \coloneqq$ the projection of the 3D motion field onto the image planes of the front and back cameras respectively.

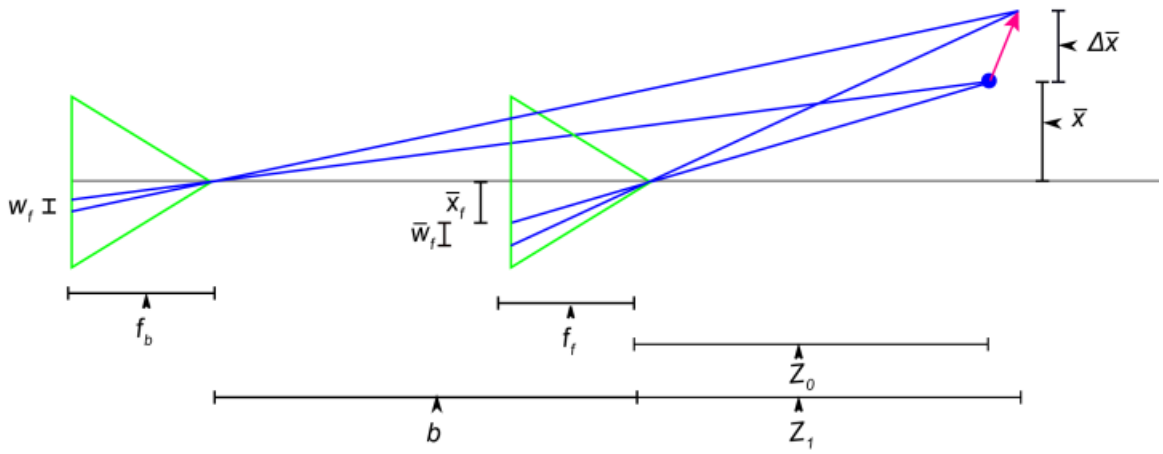


Figure 1. Coaxial camera rig geometry.

Using the projection equation to project the start point $(\bar{X}, Z_0)$ and end point $(\bar{X} + \Delta X, Z_1)$ of a point in the scene onto points in the image planes of each camera gives:

$$w_b(\bar{x}_b) = \bar{x}_{b,1} - \bar{x}_{b,0} = \frac{f_b}{(Z_1+b)}(\bar{X} + \Delta\bar{X}) - \frac{f_b}{(Z_0+b)}\bar{X} \tag{1}$$

$$w_f(\bar{x}_f) = \bar{x}_{f,1} - \bar{x}_{f,0} = \frac{f_f}{Z_1}(\bar{X} + \Delta\bar{X}) - \frac{f}{Z_0}\bar{X} \tag{2}$$

Where the second subscript of the points in the image plane represents the start or end of the projected motion.

Solving equations (1) and (2) for $\bar{X}$ and setting them equal to each other gives:

$$m(\bar{x}_f)w_f(\bar{x}_f) = c(\bar{x}_f)w_b\left(\bar{x}_f m(\bar{x}_f)\right) \tag{3}$$

where:

$$m(\bar{x}_f) = \left(\frac{f_b}{f_f}\right)\left(\frac{Z(\bar{x}_f)}{(Z(\bar{x}_f)+b)}\right) \tag{4}$$

and

$$c(\bar{x}_f) = \left(\frac{w_f(\bar{x}_f)}{\left(\frac{Z_0(\bar{x}_f)+b}{Z_1(\bar{x}_f)+b}\right)\left(\frac{Z_1(\bar{x}_f)}{Z_0(\bar{x}_f)}\right)(w_f(\bar{x}_f)+\bar{x}_f)-\bar{x}_f}\right) \tag{5}$$

$c(\bar{x}_f)$ has a direct physical interpretation. From (5), it can be seen that $c(\bar{x}_f) = 1$ if $Z_0(\bar{x}_f) = Z_1(\bar{x}_f)$ or when $\Delta Z(\bar{x}_f) = 0$. Referring to Figure 2 one can see that a change in Z introduces a slight parallax ($\rho$) in the finishing points of the optical flow detected by the two cameras. $c(\bar{x}_f)$ corrects for the parallax and can also be solved for directly from the coaxial camera geometrically.

The first term in our coaxial camera variational model is an optical flow matching term:

$$E_{match} = \int_a^b \frac{1}{2}\left[m(\bar{x}_f)w_f(\bar{x}_f) - c(\bar{x}_f)w_b\left(\bar{x}_f m(\bar{x}_f)\right)\right]^2 dx \quad (6)$$

The second term is a smoothness term:

$$E_{smooth} = \frac{1}{2}\int_a^b \left\|\nabla Z(\bar{x}_f)\right\|^2 dx \quad (7)$$

The total energy that we want to minimize is:

$$E_{total} = \gamma E_{match} + \alpha E_{smooth} \quad (8)$$

where $\gamma$ and $\alpha$ are tuning constants.

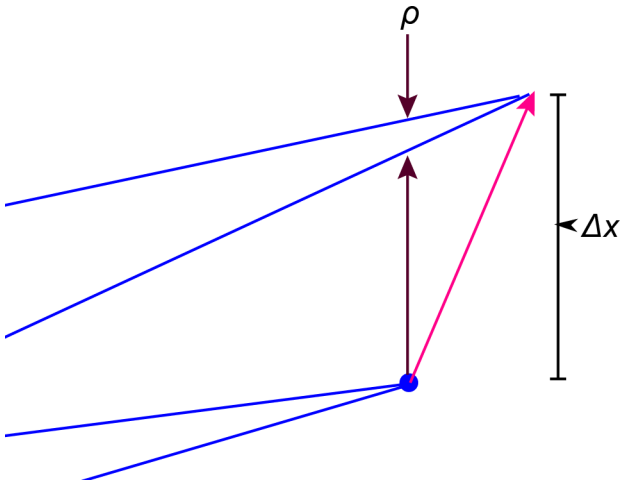


Figure 2. Parallax caused by ΔZ in a coaxial camera rig.

## 4. NUMERICAL SOLUTION

### 4.1 Euler-Lagrange

We minimize the energy $E_{total}$ by taking the Euler-Lagrange equations for (6) and (7) with respect to Z and setting to 0. The Euler Lagrange for (6) and (7) is:

$$\gamma w_z(pw_l - w_r)(m'w_f + mw_f' - c'w_b(mx) - cw_b'm_f'(mx)m'x) - \alpha\nabla^2 Z_1 = 0 \quad (9)$$

where the prime indicates the derivative with respect to Z, $\nabla^2$ is the Laplacian operator and

$$w_z = m(\bar{x}_f)w_f(\bar{x}_f) - c(\bar{x}_f)w_b\left(\bar{x}_f m(\bar{x}_f)\right) \quad (10)$$

We reduce the problem to a 1D optimization problem by observing that the solutions lie on radial epipolar lines. The Euler-Lagrange equations (one along the radial line and the other perpendicular to the radial line) are solved using the gradient descent method.

### 4.2 Initialization

We initialize the value of Z by observing that the optical flow vectors which start and end on the optical axis (e.g. $\bar{X} + \Delta\bar{X} = 0$ or $\bar{X} = 0$ ) result in a simplified version of (3) which does not depend on $\Delta Z$:

$$m(\bar{x}_f)w_f(\bar{x}_f) = w_b\left(\bar{x}_f m(\bar{x}_f)\right) \tag{11}$$

Using $Z(\bar{x}_f = (0,0)^T)$, assuming ΔZ is small relative to Z and that the scene is rigid, we use the optical flow to estimate Z for all pixels in the images. For rigid scenes with no Z translation, this is identical to the optimal solution to the Euler-Lagrange equations if the optical flow fields are equivalent to the motion fields. Where there is ΔZ and/or where the scene is not rigid, this produces a good starting point for the gradient descent iterations.

### 4.3 Resampling to a discrete grid

The gradient descent results in a new estimate of Z at $t = n + 1$ after each step, which is offset spatially in the image domain from the previous estimate of Z by the optical flow. Since optical flow algorithms produces sub-pixel flow values, the new Z values are rarely on integer pixel locations. This requires resampling the newly estimated depth map onto an integer pixel grid to obtain the Z that corresponds to each pixel.

### 4.4 Stopping Criteria

We used one of two stopping criteria depending on the quality of the flow fields and the value chosen for α. When the flow fields closely represent the motion fields and α is small (minimal Z smoothing), we use (10), which represents the mismatch in registration of the two flow fields, and stop when this number falls to the sub-pixel resolution of the optical flow algorithm being used.

Where the flow fields are noisy and not as good a representation of the motion field we needed to increase α to get good results. With more substantial smoothing, the smoothing term (7), appears to pull the Z estimate away from the correct value if γ is large and/or if many iterations are performed. In this case we stopped the iterations when the smoothing term (7) was approximately equal to, but of opposite sign to the matching term (6). This later approach produed larger residual values of $w_z$, but our experiments show that it results in more accurate depth estimations.

### 4.5 Algorithm

- Compute $\bar{w}_f$ and $\bar{w}_b$.
- Smooth $\bar{w}_f$ and $\bar{w}_b$.
- Initialize Z.
- For each radial epipolar line:
  - Iterate
    - Update Z estimate along epipolar line by updating the previous value of Z
    - Resample Z estimate to grid
  - Has stopping criteria been met?
- Repeat for next epipolar line

## 5. EXPERIMENTAL RESULTS

### 5.1 Synthetic optical flow field

For the synthetic optical flow fields we defined the geometry of a 3D scene and project the 3D motion of that scene onto a virtual image plane via an ideal pinhole camera model. This results in a simulated optical flow field that is exactly equal to the motion field. The simulated flow field experiments provide an estimate of the upper boundary of accuracy for our methodology. We determine the accuracy of the resulting image alignment by estimating the depth map along radial epipolar lines and comparing that to the original scene geometry by computing the RMS depth and disparity error.

For our synthetic flow images $f_f$ = 4.8 mm, $f_b$ = 4.0 mm, the camera has .002 mm square pixels, velocity in the XY plane was varied from 0.5 m/s to 3.5 m/s and velocity along the Z-axis ranged from 2.5 m/s toward the camera to 2.5 m/s away

from the camera. The camera frame rate was set to 30fps. We set $\gamma = 1 \cdot 10^{11}$ and $\alpha = 5 \cdot 10^{-5}$.

Figures 3 and 4 show the results for a smooth scene for a horizontal line. With the exception of the slowest XY displacement (0.5 m/s) and highest Z displacements, RMS depth error is < 0.15%. The shape of the curves suggest that there may be limitation on how large the Z displacement can be relative to the camera geometry and the XY displacement and still produce good results. We believe that this limitation may be due to cancellation which can occur between optical flow produced by lateral translation and the flow produced by forward translation. Flow due to forward translation results in radial flow and where a radial line is parallel to the direction of translational motion the flows are summed. This occurs only on one radial line and the smoothing term, which smoothes both along and between radial lines, reduces the impact.

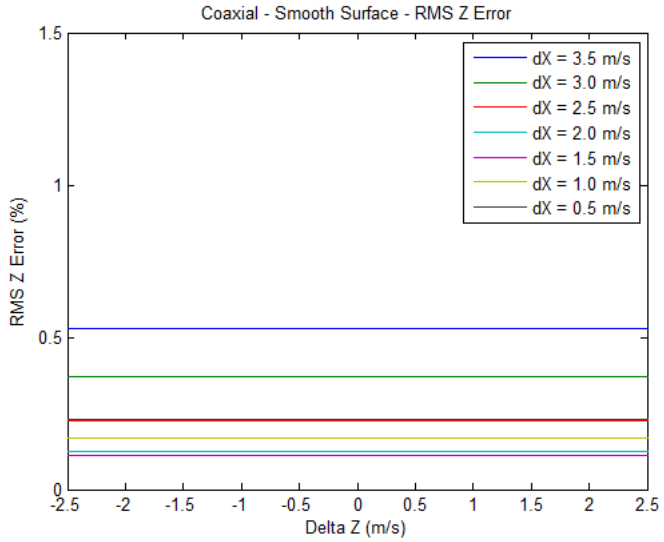


Figure 3. RMS Z error, coaxial camera rig, synthetic images, smooth surface.

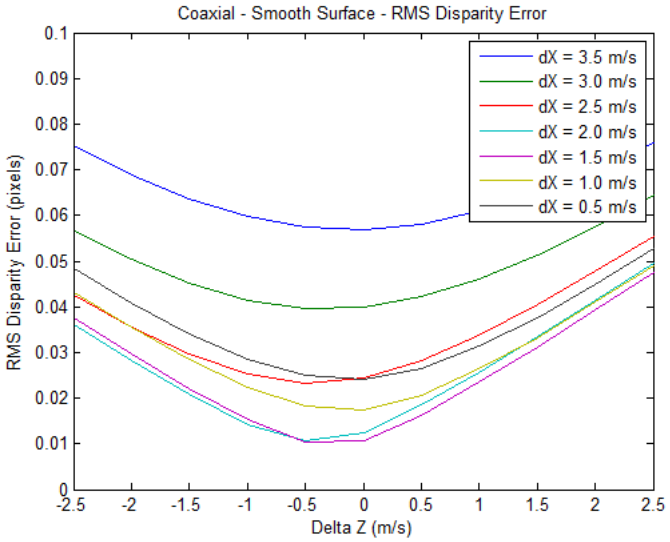


Figure 4. RMS disparity error, coaxial camera rig, synthetic images, smooth surface.

Figures 5 and 6 show the results for a scene with a large (8 m) discontinuity. As expected, the RMS error increases, but the increase is relatively minor and one would expect it to be smaller than the RMS error from comparable binocular stereo camera rig.

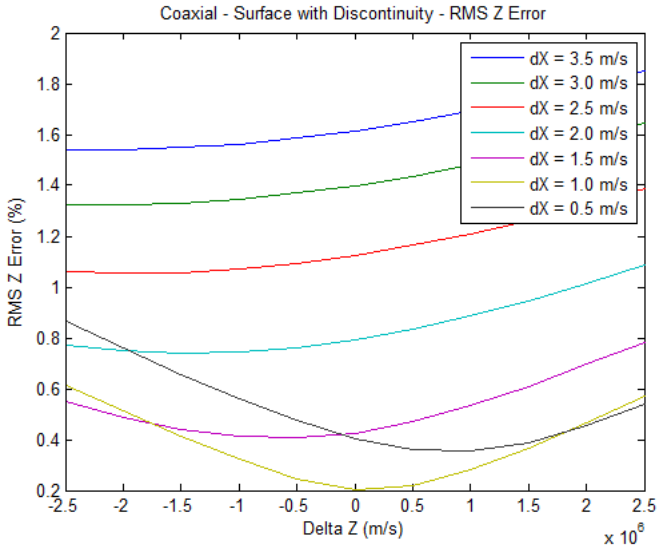


Figure 5. RMS Z error, coaxial camera rig, synthetic images, surface with discontinuities.

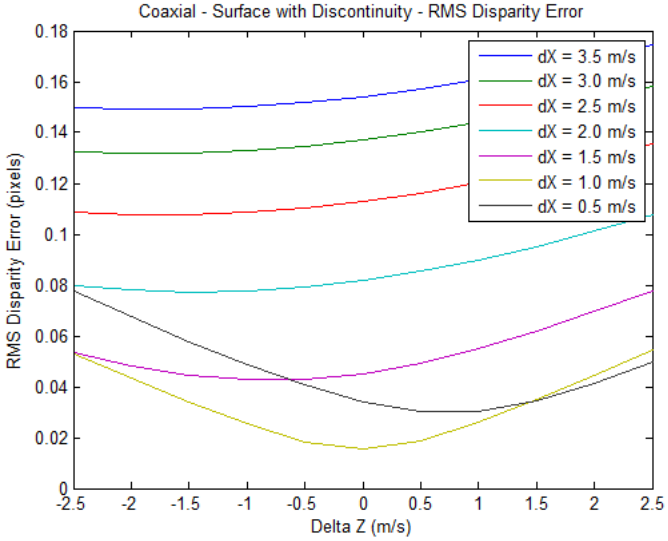


Figure 6. RMS Z disparity, coaxial camera rig, synthetic images, surface with discontinuities.

### 5.2 Real flow fields from real images

Our coaxial camera rig consists of a pair of cameras with RGB sensors mounted on a precision XY table and the camera rig was translated a known distance between frames. Accuracy was determined by comparing the estimated camera rig displacement to the known camera rig displacement and converting to disparity. A schematic representation of the camera configuration is shown in Figure 7.

Our scene, shown in Figure 8, consisted of a 10 cm diameter by 17 cm tall cylinder located 75 cm from the optical center

of the front camera in the camera rig and a planar background located 115 cm from the optical center of the front camera. There is a relatively large discontinuity between the cylinder and the planar background similar in scale to that of our synthetic optical flow experiments. Velocity in the XY plane was 0.3 m/s, which when scaled to match our synthetic optical flow fields would be 4 m/s. The cameras have 0.006 mm square pixels, focal lengths of 7.7 mm and 5.8 mm (front and back respectively), and b = 143.3 mm. We set $\gamma = 2 \cdot 10^6$ and $\alpha = .05$. We used the large scale optical flow algorithm from Brox and Malik [3]. The flow in the x and y directions were converted to radial epipolar lines at one degree increments, which provides dense reconstruction near the center of the image, but leaves some gaps near the edges of the images which we approximated by interpolation between epipolar lines when converting the depth along epipolar lines back to an XY grid.

Figure 9 shows the disparity errors. The RMS disparity error is less than 1% in the center region of the image, but at higher Z velocities and near the edges of the image where the distance between the radial lines is greatest the error increases. We believe that the increased error is primarily due to the interpolation that is required to convert the radial depth map back to an XY gird. Figure 10 shows the dense depth map after resampling onto the XY grid. In addition to the errors near the edges of the image, some anomalies due to the balance between the smoothing term along radial epipolar lines versus the smoothing term between adjacent radial epipolar lines can be seen where the horizontal support surface passes underneath the cylinder. The horizontal support surface disappears entirely from the depth map near the two bottom corners. We believe this is due partly to the larger spacing between epipolar lines in this area and partly due to the lower accuracy of the optical flow near the edges of the image.

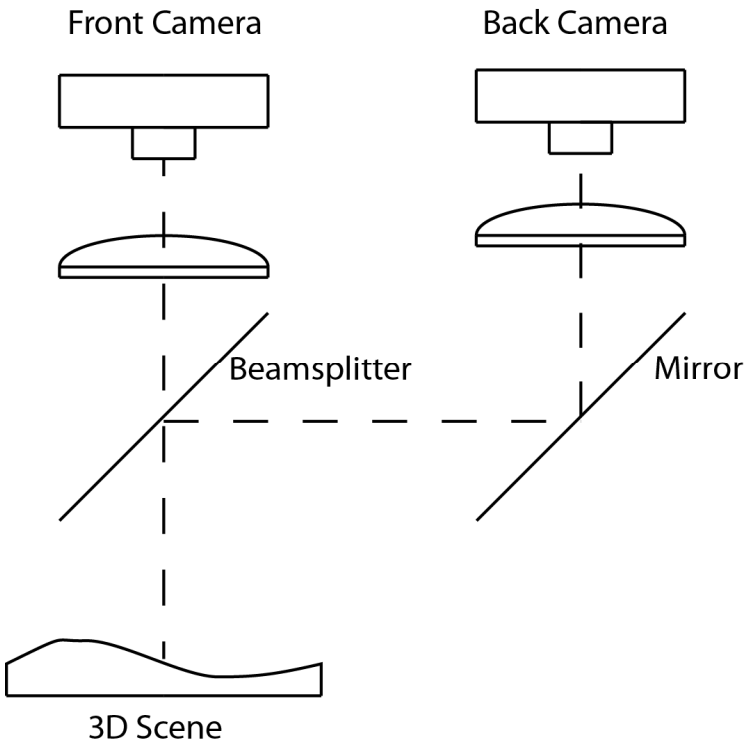


Figure 7. Coaxial camera rig.

## 6. CONCLUSIONS

Provided that there is sufficient motion between the camera and the scene and that the scene has sufficient texture to produce optical flow, our results demonstrate that depth can be estimated in the center region of a coaxial camera by aligning the optical flow fields produced by the two cameras using a variational methods approach and then estimating depth using the ratio of the flow fields. This approach overcomes the unrecoverable point problem first reported by Ma and Olsen over 25 years ago and in fact produces the most accurate results in the center of the image.

Our technique appears to be robust to flow fields that are not a good representation of the motion field as long as the flow fields in the two cameras reflect the same errors (e.g. the aperture problem and variation in illumination). This suggests that the intra-camera images might be used as an additional term in the optical flow computation (e.g. intra-camera image smoothing) to improve both the optical flow computation and the results intra-camera image alignment.

Our results also suggest that the technique could produce good results with a multimodal camera rig because optical is generally invariant to the imaged wavelength.

Figure 8. Coaxial camera rig scene.

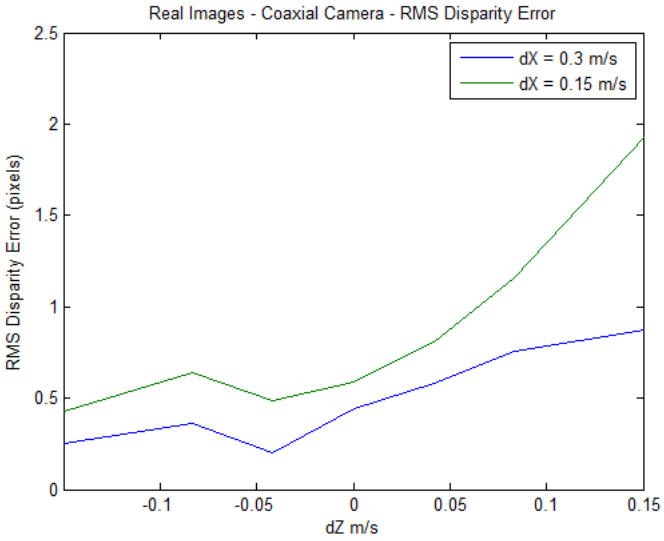


Figure 9. Disparity errors, real images, multimodal stereo camera rig.

Figure 10. Dense depth map from 3D reconstruction.

## REFERENCES

[1] N. Asada, m. Baba, and A. Oda, "Depth from Blur by Zooming," in *Proceedings of the Vision Interface Annual Conference*, Ottawa, Canada, 2001.

[2] M. Baba, N. Asada, and T. Migita, "A Thin Lens Based Camera Model for Depth Estimation from Defocus and Translation by zooming," in *Proc. 15th International Conference on Vision Interface*, Calgary, Canada, 2002.

[3] T. Brox and J. Malik, "Large Displacement Optical Flow Desriptor Matching in Variational Motion Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2010.

[4] H. Gao, J. Liu, Y. Yu, and Y. Li, "Distance measurement of zooming image for a mobile robot," *International Journal of Control, Automation and Systems,* vol. 11, pp. 782-789, 2013.

[5] A. A. Goshtasby, *Image registration principles tools methods*: Springer, 2012.

[6] R. Hartly and A. Zisserman, *Multiple View Geometry in computer vision*: Cambridge University Press, 2003.

[7] R. Kirby, "Development of a Real-Time Performance Measurement and Feedback System for Alpine Skiers," *Sports Technology,* vol. 2, pp. 43-52, 2009a.

[8] R. Kirby, "Three Dimensional Surface Mapping System Using Optical Flow US2013321790A1," USA Patent, 2012.

[9] R. Kirby and R. Whitaker, "A novel automated method for doing revistration and 3D reconstruction from multi-modal RGB/IR image sequences," in *SPIE Optics + Photonics*, San Diego, 2016.

[10] J. Lavest, G. Rives, and M. Dhome, "Three Dimensional Reconstruction by Zooming," *IEEE Transactions on Robotics and Automation,* vol. 9, pp. 196-207, 1993.

[11] J. Lavest, G. Reves, and M. Dhome, "Modeling an Object of Revolution by Zooming," *IEEE Transactions on Robotics and Automation,* vol. VOL. II, NO. 2, April 1995, 1995.

[12] J. Ma and S. I. Olsen, "Depth from Zooming," *J. Opt. Soc. Am. A* vol. 7, October 1990 1990.

[13] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two frame stereo correspondance agorithms," in *IJCV*, 2001.

[14] R. Szeliski, *Computer Vision. Algorithms and Applications*. New York: Springer, 2011.

[15] Y. Zhang and K. Qi, "Snake-Search Algorithm for Stereo Vision Reconstruction via Monocular System," presented at the The 5th Annual IEEE Conference on Cyber Technology in Automation, and Control, Intelligent Systems, Shenyang, China, 2015.

[16] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing,* vol. 21, pp. 977-1000, 2003.