3D reconstruction from images taken with a coaxial camera rig

Richard Kirby\*and Ross WhitakerUniversity 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 or borescope. 3D reconstruction from images taken through an endoscope or borescope has numerous medical and industrial applications, but until now has not found acceptance due to the lack of physical space for a traditional stereo baseline. 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, using a variational methods optimization algorithm to map pairs of optical flow fields from different focal length cameras in a coaxial camera rig. Instead of using pixel based correspondences our method uses the ratio of the optical flow fields at each pixel location to perform 3D reconstruction. This not only results in accurate image pair alignment but also produces accurate dense depth maps throughout the field of view. We test our method on synthetic optical flow fields and on real image sequences taken with a coaxial camera rig. We demonstrate our method's accuracy by evaluating our results against a ground-truth. Accuracy is comparable to a traditional binocular stereo camera rig, but without the need for the traditional stereo baseline and with substantially smaller occlusions.

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

# INTRODUCTION

3D reconstruction from image pairs taken from two different perspectives is one of the most active areas of research in computer vision [6, 11, 12]. 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, 14]. When the camera geometry is known, depth can be estimated using the disparity between the pixel pairs. However, pixels are finite in size, resulting in discrete disparity steps and discrete steps in the depth estimate. Thus the resolution of the depth estimate is a function of the camera baseline. The larger the baseline, the higher the resolution of the depth estimate. Larger baselines, however, create two well know problems: 1) The larger the baseline, the greater the occlusions (areas of the scene where one camera cannot see what the other camera sees) and 2) the larger the baseline, the larger the stereo camera rig.

There are computer vision applications where the 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/borescope or in applications where the surface being analyzed is so close to the cameras that sufficient overlap between images is impossible. 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 zero baseline camera, sometimes called depth from zooming [10] or coaxial [7] camera rig. 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 being imaged as well as 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 resolution.

In this paper we introduce a novel automated method for finding depth in image sequences taken with a coaxial camera 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.

# RElated work

Depth from images taken at different focal lengths along a common optical axis was first proposed by Ma and Olsen [10]. Lavest et al. [9, 8] 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 [13] describe a method for 3D reconstruction from multi-focal length images using a snake-search algorithm.

The original reason researchers were interested in acquiring image pairs along a common optical axes was cost, because image pairs could be acquired with a single stationary camera by changing the focal length of a zoom lens. 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 bore scope) 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 mulit-modal stereo rig.

The coaxial camera rig [7] 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. This later advantage is due to the depth estimate being derived from the ratio of the flow fields taken at different focal lengths as opposed to using the extremely small disparities found in the center region of a coaxial camera rig.

# variational model

Referring to Figure 1, let , represent points in the image domain of the front and back cameras. Let the disparity between and such that and represent the same point in the scene. Let the focal lengths for the front camera and back cameras and the distance between the optical center of the front camera and a point in the scene corresponding to , the distance being measured along the optical axis. the distance between the optical center of the two cameras. the projection of the 3D motion field onto the image planes of the front and back cameras respectively.

# Figure3

Figure 1: Coaxial camera rig geometry.

Using the projection equation to project the start point ( and end point ( of a point in the scene onto points in the image planes of each camera gives:

(1)

(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 and setting them equal to each other gives:

(3)

where:

(4)



and

(5)



has a direct physical interpretation. From (5), it can be seen that if or when . Referring to Figure 2 one can see that a change in Z introduces a slight parallax () in the finishing points of the optical flow detected by the two cameras. 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:

(6)

The second term is a smoothness term:

(7)

The total energy that we want to minimize is:

(6)

where and are tuning constants.

# Figure4

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

# Numerical Solution

## Euler-Lagrange

We minimize the energy 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:

(8)

where the prime indicates the derivative with respect to Z, is the Laplacian operator and

(9)

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 decent method.

## Initialization

We initialize the value of Z by observing that the optical flow vectors which start and end on the optical axis (e.g. or ) result in a simplified version of (3) which does not depend on :

(8)

Using , 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 fixed point iterations.

## Resampling to a discrete grid

The gradient descent results in a new estimate of Z at , 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.

## 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 equation (9), which represents the mismatch in registration of the two flow fields, and stop when this number becomes suitably small.

Where the flow fields are noisy and not as good a representation of the motion field we need to increase α to get good results. With more substantial smoothing, the smoothing term (7), can 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 (6) was approximately equal to, but of opposite sign to the matching term (7). This later approach often results in a larger error between the flow fields vs. the first approach, but our experiments show that it results in more accurate alignment because we stop iterating before the smoothness term pulls Z too far from its correct position.

## Algorithm

* Compute and .
* Smooth and .
* 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

# ExperimentAl RESULTs

## 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 and expose limitations on the 3D velocity with respect to the camera geometry. We determine the accuracy of the resulting image alignment by estimating the depth map along radial lines and comparing that to the original scene geometry by computing the RMS depth and disparity error.

For our synthetic flow images ff = 4.8 mm, fb = 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 and .

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 in certain areas in the image. 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.

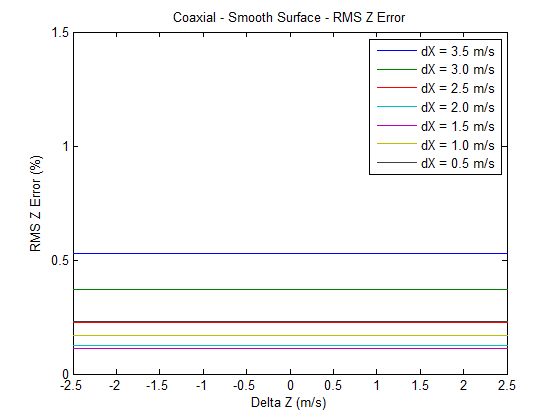


Figure 3: RMS Z error, coaxial camera rig, synthetic images, smooth surface.

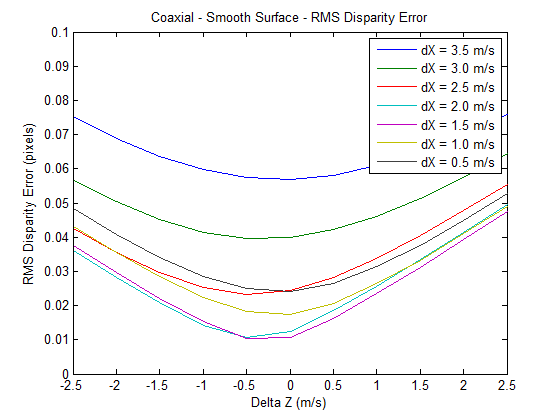


Figure 4: RMS disparity error, coaxial camera rig, synthetic images, smooth surface.

Figures 5 and 6 show the results for a scene with a large (8 m) discontinuity. The RMS error increases, but is still within acceptable levels for a wide range of applications.

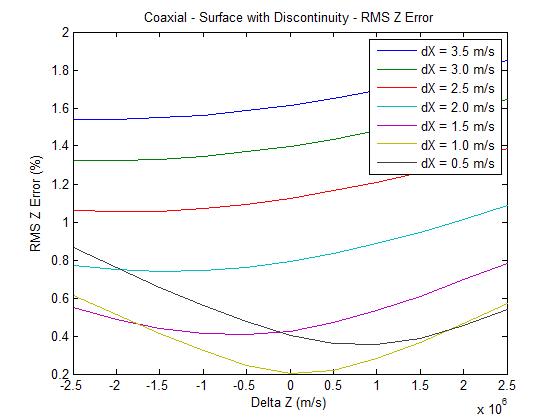


Figure 5: RMS Z error, coaxial camera rig,

synthetic images, surface with discontinuities.

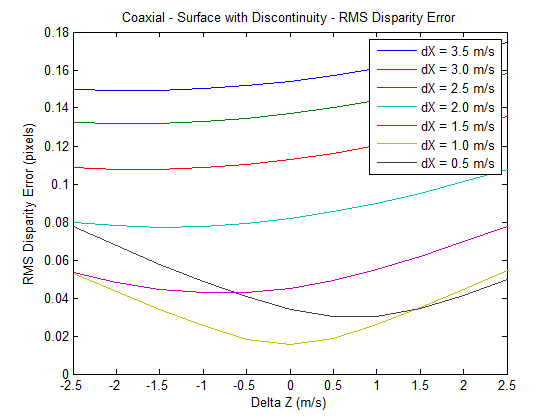


Figure 6: RMS Z disparity, coaxial camera rig,

synthetic images, surface with discontinuities.

## Real flow fields from real images

The coaxial camera rig consists of a pair of cameras with RGB sensors mounted on a precision XY table. The camera arrangement is shown in Figure 7. Depth accuracy was determined by estimating the camera movement using the estimated depth map and optical flow field and comparing the estimated camera rig displacement to the actual displacement.

Our scene (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 second set of synthetic experiments. Velocity in the XY plane was 0.3 m/s, which when scaled to match our synthetic images 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 and . We used the large scale optical flow algorithm from Brox and Malik [3].

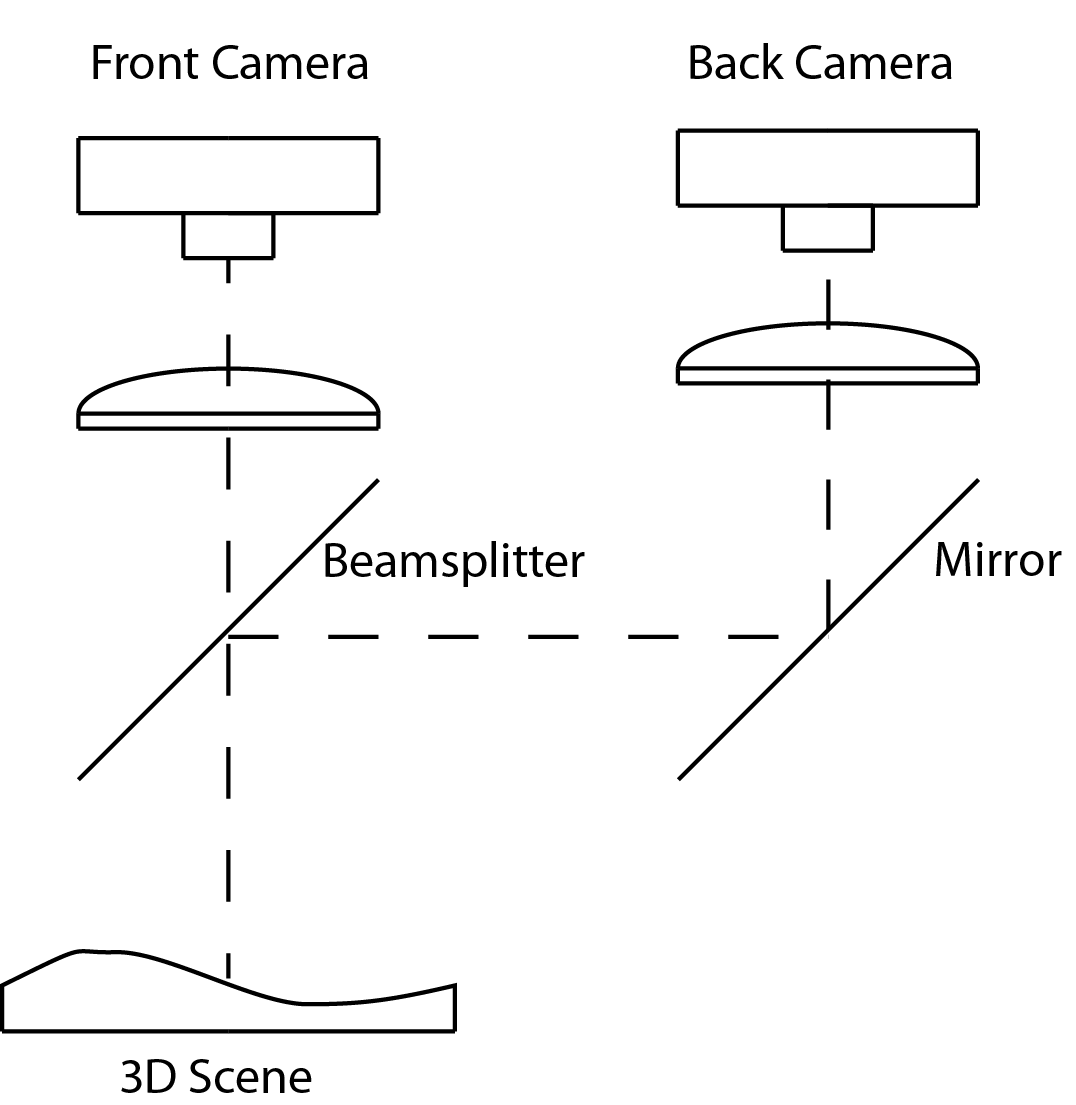


Figure 7: Coaxial camera rig..

Figure 9 shows the disparity errors. The RMS disparity error is typically less than 1% except where we get the cancellation in the flow fields between the forward translation and the lateral translation.

# Conclusions

Our results provide solid evidence that it's possible to find image correspondence using the optical flow fields provided that there is sufficient motion between the camera and the scene and that the scene has sufficient texture to produce optical flow. One advantage of our method is that images that don't have common pixel intensities or features can be aligned. Another advantage is that highly accurate sub-pixel alignment is possible in the center region of a coaxial camera. Both cases permit the estimation of dense disparity maps which can be converted into dense depth maps for 3D reconstruction and the relative velocity estimation between the scene and the camera rig.

With sufficient motion between the cameras and the scene and a scene that produces sufficient optical flow, our technique produces image alignment for a multimodal camera rig which is comparable to feature and pixel intensity based methods that align pairs of visible light images.

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 suggest that our technique could produce good results on moving multimodal camera rig (scanning security camera or vehicle mounted camera) and for a coaxial camera rig, allow stereo reconstruction in situations where a standard stereo baseline isn't feasible (e.g. endoscope or bore-scope).



Figure 8: Coaxial camera rig scene.

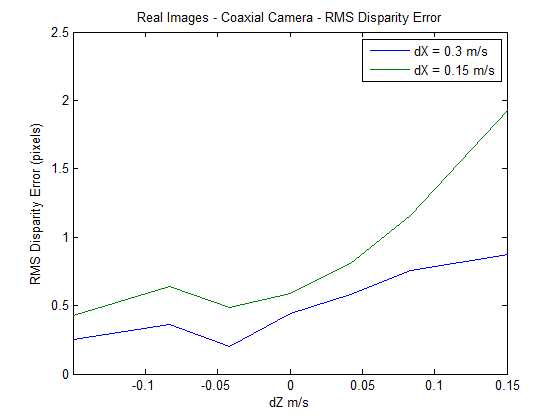


Figure9: Disparity errors, real images, multimodal stereo camera rig.

# 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, "Three Dimensional Surface Mapping System Using Optical Flow US2013321790A1," USA Patent, 2012.

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

[9] 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.

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

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

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

[13] 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.

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