A novel automated method for doing registration and 3D reconstruction from multi-modal RGB/IR image sequences

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

ABSTRACT

In recent years, the use of multi-modal camera rigs consisting of an RGB sensor and an infrared (IR) sensor have become increasingly popular for use in surveillance and robotics applications. The advantages of using multi-modal camera rigs include improved foreground/background segmentation, wider range of lighting conditions under which the system works, and richer information (e.g. visible light and heat signature) for target identification. However, the traditional computer vision method of mapping pairs of images using pixel intensities or image features is often not possible with an RGB/IR image pair. We introduce a novel method to overcome the lack of common features in RGB/IR image pairs by using a variational methods optimization algorithm to map the optical flow fields computed from different wavelength images. This results in the alignment of the flow fields, which in turn produce correspondences similar to those found in a stereo RGB/RGB camera rig using pixel intensities or image features. In addition to aligning the different wavelength images, these correspondences are used to generate dense disparity and depth maps. We obtain accuracies similar to other multi-modal image alignment methodologies as long as the scene contains sufficient depth variations, although a direct comparison is not possible because of the lack of standard image sets from moving multi-modal camera rigs. We test our method on synthetic optical flow fields and on real image sequences that we created with a multi-modal binocular stereo RGB/IR camera rig. We determine our method's accuracy by comparing against a ground truth.

Keywords: infrared, multi-modal, variational methods, optical flow, 3D, reconstruction, surveillance, binocular stereo

1. INTRODUCTION

Finding corresponding points in image pairs taken from two different perspectives is one of the most active areas of research in computer vision [7, 13, 15]. It forms the basis of 3D reconstruction as well as being a critical component of many other computer vision and image processing applications that require pixel to pixel alignment between image pairs [6, 20].

Most correspondence finding techniques are based on matching pixel intensity values or features which are derived from pixel intensity values. This, in turn, allows the estimation of dense disparity maps which, given the camera geometry, allows the estimation of dense depth maps. Where image alignment is the ultimate goal, the correspondences provide the geometrical transformation that allows one image, the sensed image, to be transformed into the second image, the reference image.

There are computer vision applications, however, where traditional correspondence finding techniques do not produce the desired results. Two notably cases are multi-modal camera rigs where the images produced from different sensor types are not similar enough to be aligned using pixel intensities or features [20] and the center region of a coaxial camera rig [12] where the disparity is too small to produce good triangulation. There are also multi-camera applications where it is desirable to augment the use of pixel intensities and/or image features to improve the finding of intra-camera correspondences.

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

In this paper we introduce a novel automated method for finding correspondences using the optical flow fields from two cameras. We apply the technique to images acquired by a multi-modal stereo rig where one camera contains an RGB sensor and the other camera contains an IR sensor. In applications where there is sufficient motion between the camera rig and the scene (scanning security camera, camera mounted on a vehicle, cameras mounted on a mobile robot, etc.), where the scene exhibits enough texture to produce optical flow, and where the scene has sufficient depth variation along epipolar lines, our method finds correspondences between multi-view image sequences without using intra-camera pixel intensities or features. From these correspondences we estimate dense disparity maps with accuracies similar to multi-modal techniques that align images based on image characteristics derived from pixel intensities.

2. RELATED WORK

Aligning images from stereo camera rigs consisting of cameras with multi-modal sensors has been an active research area for the last decade and a half. Initially inspired by the work done to match medical images to models [18] it has more recently been motivated by the need for surveillance systems that use a combination of visible light and infrared cameras to detect targets. As noted by Yaman and Kalkan [19], traditional image alignment techniques used in stereo vision are not applicable to multi-modal camera rigs because the pixel intensities can be substantially different in a visible light image vs. an IR image (Figure 1). Solutions to the multi-modal problem fall into two broad categories. The first uses Mutual Information (MI). MI was original proposed by Viola and Wells [18] to match medical images to models. Egnal [3] is considered the first to use MI as a similarity measure to match multi-modal stereo images. Since then, numerous improvements have been made including adaptive windowing [4], incorporating prior probabilities [5], regions of interest [9-11], and extending MI using gradient information [2].

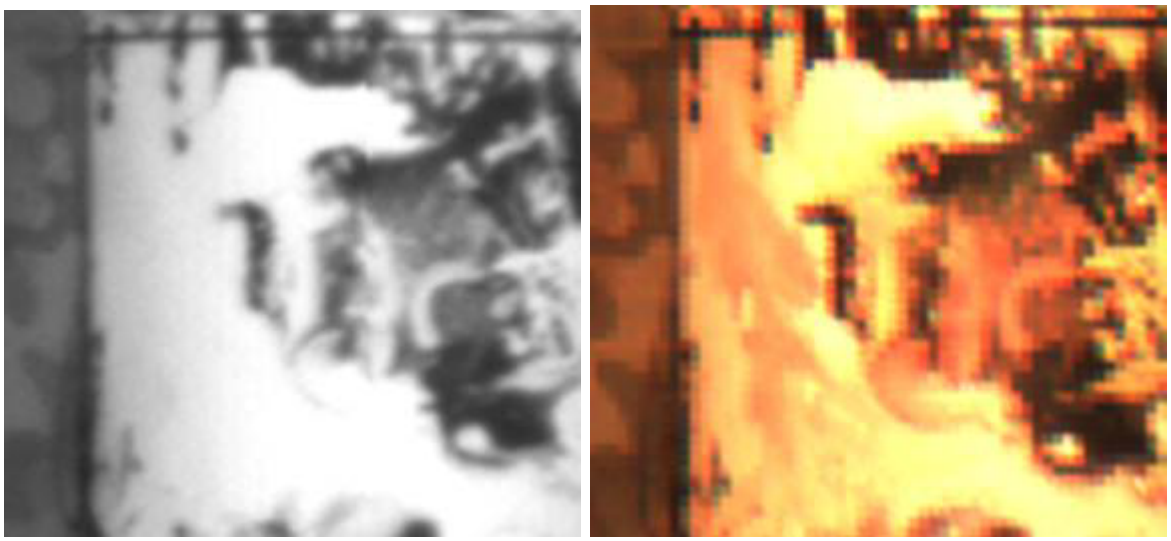

Figure 1. The left image is taken with an IR camera and the right is the same scene taken with an RGB camera.

More recently, local self-similarity (LSS), originally used in template matching, was proposed for use in a multi-modal camera rig [16]. Most recently Yaman and Kalkan [19] used MI to generate dense disparity maps from multi-modal camera rigs.

The method we present avoids using visual similarity measures between the images from the two different sensor types by computing the optical flow fields from the two sensors and then aligning the flow fields. This permits images with no visual similarity to be aligned as long as there is motion between the camera and the scene and the scene has enough texture to produce optical flow.

Verri and Poggio [17] have shown that in many cases optical flow is not equivalent to the motion field. While optical flow algorithms have improved substantially since the Verri and Poggio paper (see [14] and [1] for summaries of the progression of optical flow algorithm development); optical flow errors caused by the aperture problem, non-Lambertian surfaces, and non-uniform changing illumination, still exist.

For finding image correspondences, however, the optical flow fields do not need to be equivalent to the motion fields. For example, errors caused by the aperture problem where only the motion tangential to edges is detected or errors caused by moving shadows, will be perceived by the two sensors identically and alignment is unaffected. The primary requirement is that the optical flow computation be invariant to different light wavelengths.

3. VARIATIONAL MODEL

Referring to Figure 2, let $\bar{x}_l := (x_l, y_l)^T$, $\bar{x}_r := (x_r, y_r)^T$ represent points in the image domain of the left and right cameras. Let $\bar{h}(\bar{x}) :=$ the disparity between $\bar{x}_l$ and $\bar{x}_r$ such that $\bar{x}_l$ and $\bar{x}_r + \bar{h}(\bar{x}_r)$ represent the same point $\bar{X}(\bar{x}_l) := (X, Y)$ in the scene. Let $f :=$ the focal lengths of the cameras and $Z_0(\bar{x}_l)$, $Z_1(\bar{x}_l) :=$ the distance between the optical center of the left camera and a point in the scene corresponding to $\bar{x}_l$ at time $t = 0$ and $t = 1$, the distance being measured along the optical axis. $\Delta Z(\bar{x}_l)$ is then the difference along the Z axis for each point between $t = 0$ and $t = 1$. $\bar{X} :=$ the distance from the optical axis to a point in the scene and $\Delta\bar{X} :=$ the change in the distance from the optical axis between time $t = 0$ and $t = 1$. $b :=$ the stereo baseline. $\bar{w}_l$, $\bar{w}_r :=$ the projection of the 3D motion (the ideal flow) of point in the scene onto the image planes in the left and right cameras.

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:

$$w_l(\bar{x}_l) = \bar{x}_{l,1} - \bar{x}_{l,0} = \frac{f}{Z_1}\left(b - (\bar{X} + \Delta\bar{X})\right) - \frac{f}{Z_0}(b - \bar{X}) \tag{1}$$

$$w_r(\bar{x}_r) = \bar{x}_{l,1} - \bar{x}_{l,0} = \frac{f}{Z_1}(\bar{X} + \Delta\bar{X}) - \frac{f}{Z_0}\bar{X} \tag{2}$$

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

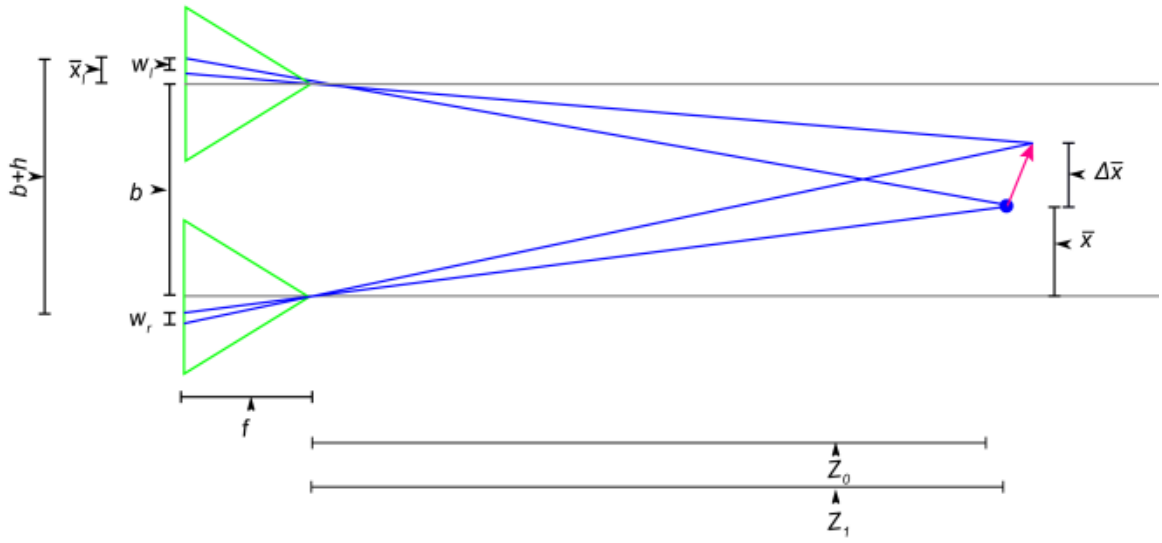

Figure 2. Binocular stereo camera rig geometry.

Solving for the ratio of the two flows $\frac{w_r(\bar{x}_l)}{w_l(\bar{x}_l + h(\bar{x}_l))}$ at corresponding points gives:

$$p(\bar{x}_l)w_l(\bar{x}_l) = w_r(\bar{x}_l + h(\bar{x}_l)) \tag{3}$$

where:

$$p(\bar{x}_l) = 1 + \frac{\Delta Z(\bar{x}_l) h(\bar{x}_l)}{f \Delta \bar{X} - \Delta Z(\bar{x}_l) \bar{x}_l} \tag{4}$$

$p(\bar{x}_l)$ has a physical interpretation. From equation (4) it can be seen that $p(\bar{x}_l) = 1$ if $\Delta Z(\bar{x}_f) = 0$. Referring to Figure 3, one can see that a change in Z introduces a slight parallax (ρ) in the finishing points of the optical flow detected by the two cameras along the image x axis. $p(\bar{x}_l)$ compensates for this parallax and can be solved for directly from the camera geometry.

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

$$E_{match} = \int_a^b \frac{1}{2} [p(\bar{x}_l) w_l(\bar{x}_l) - w_r(\bar{x}_l + h(\bar{x}_l))]^2 \, dx \tag{5}$$

The second term is a smoothness term:

$$E_{smooth} = \frac{1}{2} \int_a^b \|\nabla Z(\bar{x}_l)\|^2 \, dx \tag{6}$$

The total energy that we want to minimize is:

$$E_{total} = \gamma E_{match} + \alpha E_{smooth} \tag{7}$$

where γ and α are tuning constants.

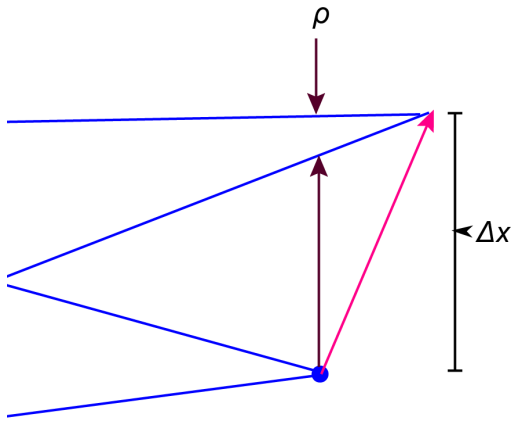

Figure 3. Parallax caused by start and end points having a different depth.

4. NUMERICAL SOLUTION

4.1 Euler-Lagrange

We minimize the energy E_{total} by taking the Euler-Lagrange equations for (5) and (6) with respect to Z and setting to 0:

$$\gamma(pw_l - w_r)(p'w_l + pw_l' - w_r'h') - \alpha \nabla^2 Z_1 = 0 \tag{8}$$

where:

$$p' = \frac{\partial p}{\partial Z} = -\frac{p}{z_1} \tag{9}$$

$$w_l' = \frac{\partial w_l}{\partial Z} = -\frac{w_l}{Z_1} \tag{10}$$

$$w_r' = \frac{\partial w_r}{\partial Z} = -\frac{w_r}{Z_1} \tag{11}$$

We reduce the problem to a 1D optimization problem by observing that the solutions for the stereo camera rig lie on horizontal epipolar lines. The Euler-Lagrange equations (one for the x direction and one for the y direction) are solved using the gradient decent method.

4.2 Initialization

We initialize the gradient decent with the value of Z from the previous image sequence. For example, image sequence 1 consists of images taken at $t = 0$ and $t = 1$; image sequence 2 consists of images taken at $t = 1$ and $t = 2$. For the first image sequence taken at $t = 0$ and $t = 1$ we use images that only consist of $\Delta\bar{X}$ translation. For robotics and surveillance applications where camera movement is controlled, this is not a significant limitation, however, one would expect the ΔZ estimate to drift over time, requiring re-initialization.

4.3 Resampling to a discrete grid

The gradient descent results in a new estimate of Z at $t = n + 1$, 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

In areas of the scene that are smooth, without depth discontinuities that generate occlusions, the flow fields computed from a stereo camera pair adhere closely to (3). However, in areas with occlusions along an epipolar line, the occlusions cause discontinuities in the flow field that violate the relationship defined in (3).

Using a stopping criteria based on the average error along an epipolar line as determined by the first term in the energy equation works well in areas without occlusions, but distorts the depth estimation in occluded areas. Additionally, running the gradient descent too long causes the smoothing term (6) in the energy equation to propagate errors caused by occlusion to adjacent pixels, reducing the overall accuracy of the depth estimation.

To solve this stopping criteria problem, we propose using optical flow as an occlusion detector. When there is a step function in the optical flow field along the axis parallel to the camera rig baseline, there will be an occlusion. Occlusions are thus detectable and the occluded areas can be removed from the computation of the average error along an epipolar line. This method stops the gradient decent when average error in non-occluded areas reaches a sufficiently small value. Even with the above approach, there will be some smoothing caused by the second term in the energy equation that propagates to pixels adjacent to discontinuities. An area for future exploration would be to use a discontinuity preserving smoothing term or a piecewise optimization.

4.5 Algorithm

- Compute $\bar{w}_l$ and $\bar{w}_r$.
- Smooth $\bar{w}_l$ and $\bar{w}_r$.
- Initialize Z.
- For each 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

While continually improving, in many situations today's optical flow algorithms do not produce flow fields that are equivalent to the projection of the motion field onto the camera image plane [17]. Some common flow field errors, like those created by the aperture problem, produce identical errors in both cameras in a stereo camera rig, some do not and the ones that do not will reduce the accuracy of our method. In order to determine the upper limit of the accuracy, we first tested the algorithm on synthetic flow fields that are a projection of the motion field. As optical flow algorithms continue to improve and produce results that are better approximations of the projected motion field, the results achieved with synthetic images will be approached.

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 identical to the projected motion field.

To determine the accuracy of the resulting image alignment we reconstruct the depth map along horizontal epipolar lines using the results of registration and compare the reconstructed depth map with the original scene geometry computing both the RMS disparity error and the resulting RMS depth error.

For our synthetic flow images we created a scene geometry that ranges from 10 m to 20 m from the camera center. f = 4.0 mm, the cameras have .006 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 30 fps. We set $\gamma = 1 \cdot 10^{9}$ and $\alpha = 1 \cdot 10^{-1}$.

Figures 4 and 5 show the results for a smooth scene without occlusions. The worst case RMS depth error is < 0.25% and worst case RMS disparity errors < 0.01 pixels. The accuracy is slightly reduced as delta Z increases and delta X decreases. We believe that this slight reduction in accuracy is due to the cancellation that occurs in the flow fields between X and Z translations in some areas of the image.

Figures 6 and 7 show the results for a scene with a large occlusion caused by a large (8 m) discontinuity. The RMS error increases slightly.

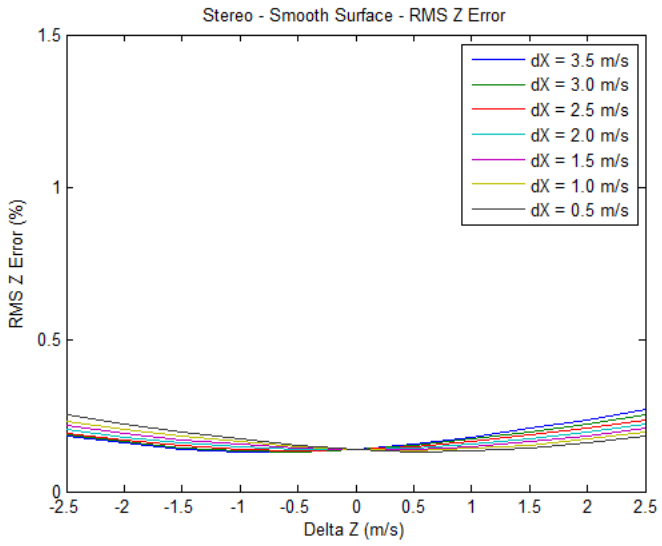

Figure 4. RMS Z error, synthetic flow field, no occlusions.

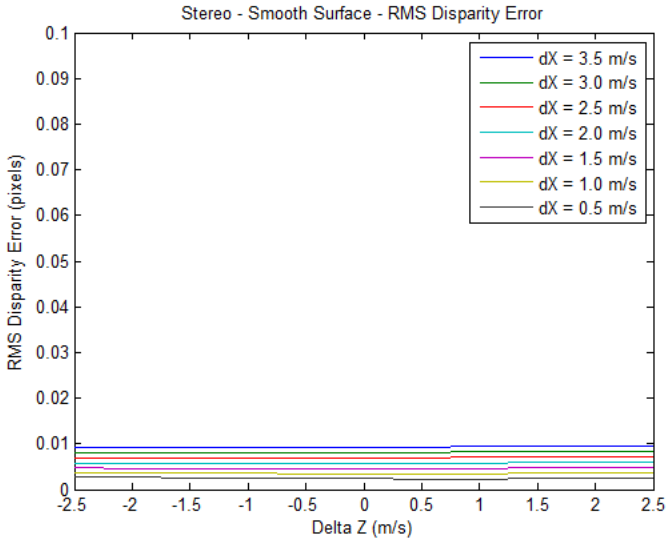

Figure 5. RMS disparity error, synthetic flow field, no occlusions.

5.2 Real flow fields from real images

Our multi-modal stereo camera rig consists of one camera with an RGB sensor and a second camera with an IR sensor. The camera rig was 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.

Our scene is shown in Figure 8. There are occlusions between the geometric shapes and between the geometric shapes and the planar background. Velocity in the XY plane was varied between 0.15 m/s and 0.3 m/s, which when scaled to match our synthetic images would be about 4 m/s. The cameras in the stereo rig had 5.3 micron (IR) and 6 micron (RGB) square pixels and 7.0 mm (IR) and 7.7 mm (RGB) focal lengths. The images were rectified to correct for the difference in pixel size and focal length. The baseline b = 75 mm. Gamma ranged from 0.2 to 0.5 and alpha was set at 0.01. To compute optical flow, we used the large scale optical flow algorithm from Brox and Malik [1].

Figure 9 shows the average disparity errors for the entire image, however, the disparity errors are not evenly distributed across the image. Disparity errors are typically less than 1 pixel where there are depth variations along epipolar lines (e.g. where the epipolar lines are passing through the geometric objects in the foreground). Conversely, on frontal-planar surfaces, like the background, where there is no depth variation along segments of epipolar lines, optical flow is a constant value across a series of pixels. For an algorithm that finds correspondences by matching two flow fields, a constant flow region is the equivalent of a featureless region for feature based correspondence finding algorithms.

From equation (5), it can be seen that for constant flow, any value of $h(\bar{x}_l)$ that maps any part of a constant flow field from one camera to any part of the corresponding constant flow field from the other camera will result in the same matching energy. From equation (6) we see that the smoothness term will favor the disparity of the closest neighboring pixel from the non-constant flow part of the flow field. This is particularly problematic for estimating the disparity of frontal planar regions separated from non-frontal planar regions by a discontinuity. Where this occurs (as in our test scene) the algorithm performs well in estimating the disparity of regions that have depth variations, but performs poorly in estimating the disparity of the regions without depth variations. The equivalent situation exists for traditional stereo correspondence finding methods that match image features. For feature matching algorithms, if a featureless region is separated from a region with features by a discontinuity, the disparity at the last matched feature will propagate into the featureless region if a smoothing term is used.

Unlike methods that use image features, however, there is a solution for this problem when aligning optical flow fields.

A camera rig where each camera has a different focal length optical system will produce flow fields that are proportional in magnitude to the ratio of the focal lengths. This characteristic can be used to accurately align constant flow regions based on the ratio of the flow magnitude from the two cameras. We present the mathematical derivation and provide a demonstration of the effectiveness of using multi-focal length optics in a system that estimates depth from flow field alignment using a coaxial camera rig in [8]. The extension of the work we present in this paper to a binocular stereo camera rig with two different focal length optical systems will be the focus of future research.

6. CONCLUSIONS

Our results provide evidence that it's possible to find image correspondences using the optical flow fields from different mode cameras provided that there is sufficient motion between the camera and the scene, that the scene has sufficient texture to produce optical flow, and that the scene has sufficient depth variation along epipolar lines. One advantage of our method is that images that don't have common pixel intensities or features can be aligned. This permits the estimation of dense disparity maps which can be converted into dense depth maps for 3D reconstruction and relative velocity estimation between the scene and the camera rig.

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 optical flow computation (e.g. add an energy term for intra-camera flow field consistency) to improve both the accuracy of the optical flow computation and the resulting intra-camera image alignment.

Our results also suggest that our technique could produce good results on a moving multi-modal camera rig (scanning security camera or vehicle mounted camera) where an initialization procedure can be performed with only translation parallel to the image plane.

With a binocular stereo camera rig in which each camera has the same focal length optical system, our method does not perform well on images with small variations in depth or were one region has depth variations but is separated by a discontinuity from another region without depth variation. We believe this can be overcome by incorporating different focal length imaging optics into the camera rig, which will be the focus of our future research.

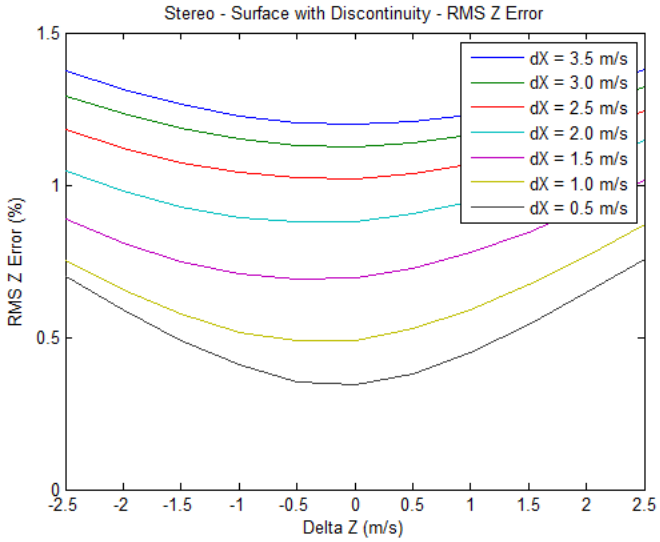

Figure 6. RMS Z error, multi-modal stereo rig, synthetic images, surface with discontinuities.

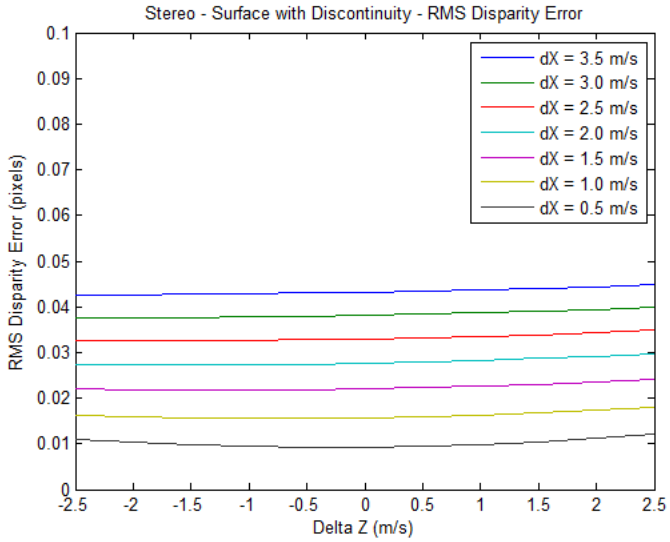

Figure 7. RMS disparity error, multi-modal stereo rig, synthetic images, surface with discontinuities.

Figure 8. Multi-modal stereo rig scene.

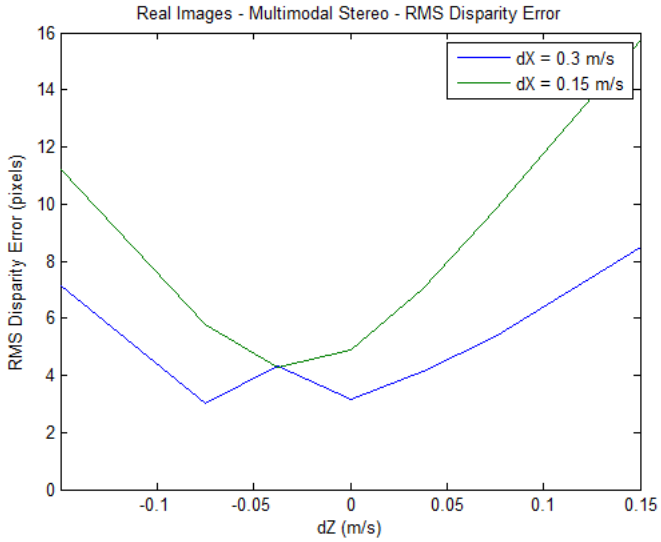

Figure 9. Disparity errors, real images, multi-modal stereo camera rig.

REFERENCES

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

[2] F. B. Campo, R. L. Ruiz, and A. D. Sappa, "Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation," *IEEE Journal of Selected Topics In Signal Processing, vol. 6, no.5,* 2012.

[3] G. Egnal, "Mutual information as a Stereo Correspondence Measure," *Technical Report MS-CIS-00-20, Computer and Information Science, University of Pennsylvania, Philadelphia, USA,* 2000.

[4] C. Fookes, A. Lamanna, and M. Bennamoun, "A new stereo image matching technique using mutual information," in *In Proceedings of the International Conference on Computer, Graphics and Imaging, CGIM'01, pages 168-173, Honolulu, USA, 2001. Iasted, ISBN 0-88986-303-2.*, 2001.

[5] C. Fookes, S. Maeder, S. Sridharan, and J. Cook, "Multi-Spectral Stereo Image Matching using Mutual Information," in *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, 2004.

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

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

[8] R. kirby and R. Whitaker, "3D reconstruction from images taken with a coaxial camera rig " in *SPIE Optics + Photonics*, San Diego, 2016.

[9] S. Krotosky and T. Mohan, "Registration of Multimodal Stereo Images using Disparity Voting from Correspondence Windows," in *IEEE Conf. on Advanced Video and Signal based Surveillance (AVSS'06)*, 2006.

[10] S. Krotosky and M. Trivedi, "Multimodal Stereo Image Registration for Pedestrian Detection," in *in Proc. IEEE Intell. Transp. Syst. Conf., Sep. 2006, pp. 109–114*, 2006.

[11] S. Krotosky and M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Computer Vision and Image Understanding,* vol. 106, pp. 270-287, 2007.

[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] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two frame stereo correspondence algorithms," *International Journal of Computer Vision* vol. 47, pp. 7-42, 2002.

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

[16] A. Toraby and G. Bilodeau, "Local self-similarity as a dense stereo correspondence measure for themal visible video registration," *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on,* pp. 61 - 67, 2011.

[17] A. Verri and T. Poggio, "Motion Field and Optical Flow: Qualitative Properties," *IEEE Trans. Pattern Analysis and Machine Intelligence, 11(5), pp. 490{498, May,* 1989.

[18] P. Viola and W. M. Wells, "Alignment by Maximization of Mutual Information," *Intl. J. of Computer Vision, vol. 24, no. 2, pp. 137–154,,* 1997.

[19] M. Yaman and S. Kalkan, "An iterative adaptive multi-modal stereo-vision method using mutual information," *Journal of Visual Communication and Image Representation,* vol. 26, pp. 115-131, 2015.

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