Gradient Based Multi-modal Sensor Calibration

Zachary Taylor and Juan Nieto
University of Sydney, Australia
{z.taylor, j.nieto}@acfr.usyd.edu.au

***Abstract*— This paper presents an evaluation of a new metric for registering two sensors of different modality. The metric operates by aligning gradients present in the two sensors' outputs. This metric is used to find the parameters between the sensors that minimizes the misalignment of the gradients. The metric can be applied to a wide range of problems and has been successfully demonstrated on the extrinsic calibration of two different lidar-camera systems as well as the alignment of IR and RGB images. Unlike most of previous techniques, our method requires no markers to be placed in the scene and can operate on a single scan from each sensor.**

I. Introduction

Most mobile robotic platforms rely on a large range of different sensors to navigate and understand their environment. However before multiple sensors can work together to give information on the same target the sensor outputs must be registered. This registration is far from trivial due to the very different modalities via which different sensors may operate. This registration has traditionally been performed by either hand labelling points or placing markers such as corner reflectors or chequerboards in the scene. The location of these markers are detected by all of the sensors and their positions are used for calibration.

The calibration produced by hand-labelling or maker-based methods, while initially accurate, is quickly degraded due to the robot's motion. For mobile robots working on topologically variable environments, such as agricultural or mining robots, the motion can result in significantly degraded calibration after as little as a few hours of operation. Under these conditions marker based calibration quickly becomes tedious and impractical. To maintain an accurate calibration, an automated system that can recalibrate the sensors using observations made during the robot's normal operations is required. We envision a system that would periodically retrieve a set of scans from the sensors and then, while the robot continues its tasks, process it to validate the current calibration and update the parameters when needed.

Towards that aim, we have developed a new metric, the *gradient orientation measure* (GOM) that can effectively align the outputs of two sensors of different modalities. The metric can calibrate multi-sensor platforms by optimising through a set of observations, and, unlike most current calibration approaches, the metric is also able to calibrate from a single scan pair. This last property makes our approach suitable for a broad range of applications since it is not restricted to calibration based on multiple observations from sensors attached to a rigid mount. To demonstrate the metric's potential and versatility we present results on three different datasets: (i) the alignment of two hyper-spectral camera images, (ii) the calibration of a rotating panoramic camera with a single high resolution scan and (iii) the calibration of a panospheric camera with a series of Velodyne scans. In each of these tests the proposed approach is compared with state of the art methods. An example of the results obtained with our system is shown in Figure 1.

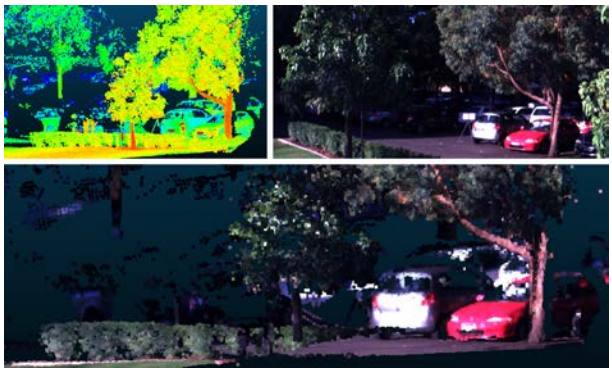


Fig. 1. Camera and lidar scan being combined. Raw lidar data is shown in the top left, with the camera image shown in the top right. The textured map obtained with our approach is shown at the bottom.

II. Related work

The most common techniques in multimodal registration are *mutual information* (MI) and *normalized mutual information* (NMI). Both measures use Shannon entropy to give an indication of how much one sensor output depends on the other. They have both been widely used in medical image registration, a survey of MI-based techniques has been presented in [1].

A. Mastin et al. achieved registration of an aerial lidar scan by creating an image from it using a camera model [2]. The intensity of the pixels in the image generated from the lidar scan was either the intensity of the laser return or the height from the ground. The images were compared using the joint entropy of the images and optimisation was done via downhill simplex. The method was only tested in an urban environment where buildings

provided a strong relationship between height and image colour.

One of the first approaches used to successfully register Velodyne scans with camera images that did not rely on markers was presented in [3]. Their method operates on the principle that depth discontinuities detected by the lidar will tend to lie on edges in the image. Depth discontinuities are isolated by measuring the difference between successive lidar points and removing points with a depth change of less than 30 cm. An edge image is produced from the camera that is then blurred to increase the capture region of the optimiser. The average of all of the edge images is then subtracted from each individual edge image to remove any bias to a region. The two outputs are combined by projecting the isolated lidar points onto the edge image and multiplying the magnitude of each depth discontinuity by the intensity of the edge image at that point. The sum of the result is taken and a grid search used to find the parameters that maximise the resulting metric.

Two very similar methods that also operate on Velodyne-camera systems have been independently developed by Pandey et al. [4] and Wang et al. [5]. These methods use the known intrinsic values of the camera and estimated extrinsic parameters to project the lidar's scan onto the camera's image. The MI value is then taken between the lidar's intensity of return and the intensity of the corresponding points in the camera's image. When the MI value is maximised, the system is assumed to be perfectly calibrated. The only major difference between these two approaches is in the method of optimisation used; Pandey et al. makes use of the Barzilai-Borwein (BB) steepest gradient ascent algorithm, while R. Wang et al. makes use of the Nelder-Mead downhill simplex method. In both implementations, aggregation of a large set of scans is required for the optimisers used to converge to the global maximum.

III. Multi-modal sensor calibration

Our method can be divided into two main stages: feature computation and optimisation.

The feature computation stage converts the sensor data into a form that facilitates comparisons of different alignments during the optimisation stage. The initial step is to perform histogram equalisation on the input intensities to ensure high contrast in the data. Next, an edge detector is applied to the data to estimate the intensity and orientation of edges at each point; the edge detector used depends on the dimensionality of the data. The strength of the edges is histogram equalised to ensure that a significant number of strong edges are present. This edge information is finally passed into the optimisation, completing the feature computation step.

The sensors' outputs are aligned during the optimisation. This is done by defining one sensor's output as fixed (called the base sensor output) and transforming the other sensor's output (referred to as the relative sensor output). In our framework, the base output is always 2D. For two 2D images, an affine transform is used, and for 2D-3D alignment, a camera transform is used to project the 3D points of the relative output onto the 2D base output. Once this has been done, the base output is interpolated at the locations that the relative output was projected onto to give the edge magnitudes and directions at these points.

Finally, GOM is used to compare the edge features between the two outputs and to provide a measure of the quality of the alignment. This process is repeated for different transformations until the optimal set of parameters is found. [1]

A. Transformation

The transformation applied to align the sensors' outputs depends on the dimensionality of the two sensors. If one sensor outputs 3D data, for example a lidar, and the other sensor is a camera, then a camera model is used to transform the 3D output. If both sensors provide a dense 2D image, then an affine transform is used to align them. A more detailed look at calculating the transforms is covered in [6].

B. Gradient calculation

The magnitude and orientation of the gradient of a camera's image intensity is calculated using the Sobel operator. Calculation of the gradient from 3D data sources is slightly more challenging and performed using the method outlined in [6]

C. The Gradient orientation measure

The formation of a measure of alignment between two multi-modal sources is a challenging problem. Strong features in one source can be weak or missing in the other. A reasonable assumption when comparing two multi-modal images is that, if there is a significant change in intensity between two points in one image, then there is a high probability there will be a large change in intensity in the other modality. This correlation exists as these large changes in intensity usually occur due to a difference in the material or objects being detected.

GOM exploits these differences to give a measure of the alignment. GOM operates by calculating how well the orientation of the gradients are aligned between two images. For each pixel, it gives a measure of how aligned the points are by taking the absolute value of the dot product of the gradient vectors:

$$alignment_j = |g_{(1,j)} \cdot g_{(2,j)}| \tag{1}$$

where $g_{(i,j)}$ is the gradient in image i at point j. The absolute value is taken, as a change going from low to high intensity in one modality may be detected as going from high to low intensity in the other modality.

[1] All the code used for our method as well as additional results and documentation is publicly available online at http://www.zacharyjeremytaylor.com

Summing the value of these points results in a measure that is dependent on the alignment of the gradients. An issue, however, is that this measure will favour maximising the strength of the gradients present in the overlapping regions of the sensor fields. To correct for this bias, the measure is normalised after the sum of the alignments has been made, by dividing by the sum of all of the gradient magnitudes. This gives the final measure as shown in Equation 2.

$$GOM = \frac{\sum_{j=1}^{n} |g_{(1,j)} \cdot g_{(2,j)}|}{\sum_{j=1}^{n} \|g_{(1,j)}\| \|g_{(2,j)}\|} \quad (2)$$

The measure has a range from 0 to 1, where, if 0, every gradient in one image is perpendicular to that in the other, and 1 if every gradient is perfectly aligned. Some typical GOM values for a range of images is shown in Figure 2. The NMI values are also shown for comparison.

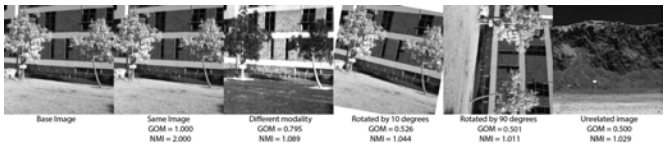


Fig. 2. GOM and NMI values when the base image shown on the left is compared with a range of other images.

D. Optimisation

The registration of one-off scans, and the calibration of a multi-sensor system tend to have significantly different constraints on their optimisation. Because of this, our approach for optimising each problem differs.

For cases where multiple scans can be aggregated, the optimisation is performed using the Nelder-Mead simplex method [7] in combination with a Gaussian pyramid. In our experiments, four layers were used in the pyramid, with Gaussians with σ of 4, 2, 1 and 0 applied.

When optimization from a single scan is required and/or there is significant error in the initial guess for the calibration, the search space becomes highly non-convex and a local optimization method such as Nelder-Mead cannot reliably find the global minimum. In these situations the metric is optimized using Particle swarm. Particle swarm optimisation works by randomly placing an initial population of particles in the search space. On each iteration a particle moves to a new location chosen by summing three factors: i) it moves towards the best location found by any particle, ii) it moves towards the best location it has ever found itself and iii) it moves in a random direction. The optimiser stops once all particles have converged. The implementation of particle swarm used was developed by S Chen [8]. In our experiments we used a particle swarm optimiser with 500 particles.

IV. Experimental Results

A. Metrics Evaluated

In this section, a series of metrics are evaluated on three different datasets. The metrics evaluated are as follows:

- MI - mutual information, the metric used by Pandey et al. [4] in their experiments on the Ford dataset [9].
- NMI - normalised mutual information, a metric we had used in our previous work on multi-modal calibration [10].
- The Levinson method [3].
- GOM - the gradient orientation measure developed in this paper.
- SIFT - scale invariant feature transform, a mono-modal registration technique included to highlight some of the challenges of multi-modal registration and calibration.

B. Parameter Optimisation

To initialise the optimisation we use either the ground truth (when available) or a manually calibrated solution. We then added a random offset to it. The random offset is uniformly distributed, with the maximum value used given in the details of each experiment. This random offset is introduced to ensure that the results obtained from multiple runs of the optimisation are a fair representation of the method's ability to converge to a solution reliably. When particle swarm optimisation is used, the search space of the optimiser is set to be twice the size of the maximum offset.

On datasets where no ground truth was available the search space was always constructed so that the space was much greater than twice the estimated error of the manual calibration to ensure that it would always be possible for a run to converge to the correct solution. All experiments were run 10 times with the mean and standard deviation from these runs reported for each dataset.

C. Dataset I

A Specim hyper-spectral camera and Riegl VZ1000 lidar scanner were mounted on top of a Toyota Hilux and used to take a series of four scans of our building, the Australian Centre for Field Robotics from the grass courtyard next to it. The focal length of the hyper-spectral camera was adjusted between each scan. This was done due to the different lighting conditions and to simulate the actual data collection process in the field.

This dataset required the estimation of an intrinsic parameter of the camera, its focal length in addition to its extrinsic calibration. To test the robustness and convergence of the methods, each scan was first roughly manually aligned. The search space was then constructed assuming the roll, pitch and yaw of the camera were each within 5 degrees of the lasers. The camera's principal distance was within 40 pixels of correct (for this camera

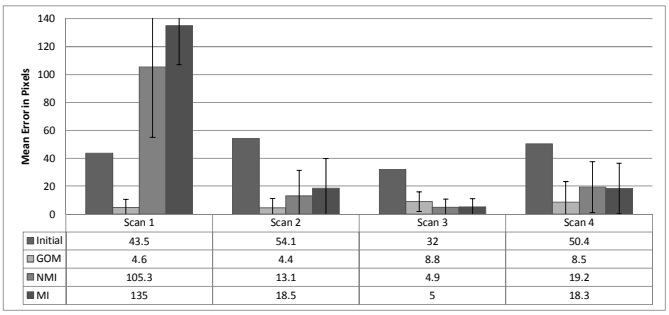


TABLE I
ACCURACY COMPARISON OF DIFFERENT METHODS ON ACFR DATASET. ALL DISTANCES IN PIXELS

principal distance ≈ 780) and the X, Y and Z coordinates were within 1 metre of correct.

1) Results: No accurate ground truth is available for this dataset. To overcome this issue and allow an evaluation of the accuracy of the method, 20 points in each scan-image pair were matched by hand. An evaluation of the accuracy of the method was made by measuring the distance in pixels between these points on the generated images. The results are shown in Table I.

For this dataset GOM significantly improved upon the initial guess for all four of the tested scans. Scans 1 and 2 were however more accurately registered then scans 3 and 4. These last two scans were taken near sunset, and the long shadows and poorer light may have played a part in the reduced accuracy of the registration. NMI gave mixed results on this dataset, misregistering scan 1 by a large margin and giving results far worse than GOM's for scans 2 and 4. It did however outperform all other methods on Scan 3. MI gave a slightly worse, but similar, performance. Levinson's method could not be evaluated on this dataset as it requires multiple images to operate.

D. Dataset II

To test each method's ability to register different modality camera images such as IR-RGB camera alignment, two scenes were scanned with a hyper-spectral camera. Bands near the upper and lower limits of the camera's spectral-sensitivity were selected so that the modality of the images compared would be as different as possible, providing a challenging dataset on which to perform the alignment. The bands selected were at 420 nm (violet light) and 950 nm (near IR). The camera was used to take a series of three images of the ACFR building and three images of cliffs at a mine site. An example of the images taken is shown in Figure 3.

The search space for the particle swarm optimiser was setup assuming the X and Y translation were within 20 pixels of the actual image, the rotation was within 10 degrees of the actual image, the X and Y scale were within 10 % of the actual image and the x and y shear were within 10 % of the actual image.

Fig. 3. Images captured by hyper-spectral camera. The top image was taken at 420nm and the bottom at 950nm

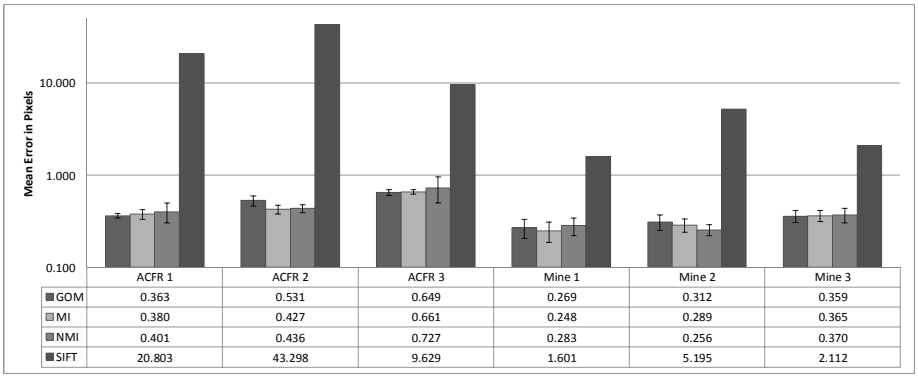
1) Results: In addition to the GOM, MI and NMI methods that have been applied to all of the datasets, SIFT features were also used. SIFT was used in combination with RANSAC to give the final transform. To measure how accurate the registration was, the average difference in position between each pixel's transformed position and its correct location was obtained. The results of this registration are shown in Table II. The images taken at the ACFR were 320 by 2010 pixels in size. The width of the images taken at the mine varied slightly, but were generally around 320 by 2500 pixels in size.

SIFT performed rather poorly on the ACFR dataset and reasonably on the mining dataset. The reason for this difference was most likely due to the very different appearance vegetation has at each of the frequencies tested. This difference in appearance breaks the assumption SIFT makes of only linear intensity changes between images, and therefore the grass and trees at the ACFR generate large numbers of incorrect SIFT matches. In the mine sites that are devoid of vegetation, most of the scene appears very similar, allowing the SIFT method to operate and give more accurate results.

Looking at the mean values for each run MI, NMI and GOM gave similar performance on these datasets, all achieving sub-pixel accuracy in all cases. There was little variation in the results obtained using the multi-modal metrics, with all three methods always giving errors between 0.2 and 0.8 pixels.

E. Dataset III

The Ford campus vision and lidar dataset has been published by G. Pandey et al. [9]. The test rig was a Ford F-250 pick-up truck which had a Ladybug panospheric camera and Velodyne lidar mounted on top. The dataset contains scans obtained by driving around downtown Dearborn, Michigan USA. An example of the data is shown in Figure 4. The methods were tested on a subset of 20 scans. These scans were chosen as they were the same scans used in the results presented by Pandey et al. Similarly, the initial parameters used were those provided with the dataset. As all of the scan-image pairs on this dataset shared the same calibration parameters, aggregation of the scans could be used to improve the accuracy of the metrics. Because of this, each experiment was performed three times, aggregating 2, 5 and 10 scans.



	ACFR 1	ACFR 2	ACFR 3	Mine 1	Mine 2	Mine 3
GOM	0.363	0.531	0.649	0.269	0.312	0.359
MI	0.380	0.427	0.661	0.248	0.289	0.365
NMI	0.401	0.436	0.727	0.283	0.256	0.370
SIFT	20.803	43.298	9.629	1.601	5.195	2.112

TABLE II

Error and standard deviation of different registration methods performed on hyperspectral images. Error is given as the mean per-pixel-error in position. Note that the chart's axis uses a log scale.

Fig. 4. Overlapping region of camera image (top) and lidar scan (bottom) for a typical scan-image pair in the Ford Dataset. The lidar image is coloured by intensity of laser return

1) Results: The Ford dataset does not have a ground truth. However, a measure of the calibration accuracy can still be obtained through the use of the Ladybug camera. The Ladybug consists of five different cameras all pointing in different directions (excluding the camera pointed directly upwards). The extrinsic location and orientation of each of these cameras is known very accurately with respect to one another. This means that if the calibration is performed for each camera independently, the error in their relative location and orientation will give a strong indication as to the method's accuracy.

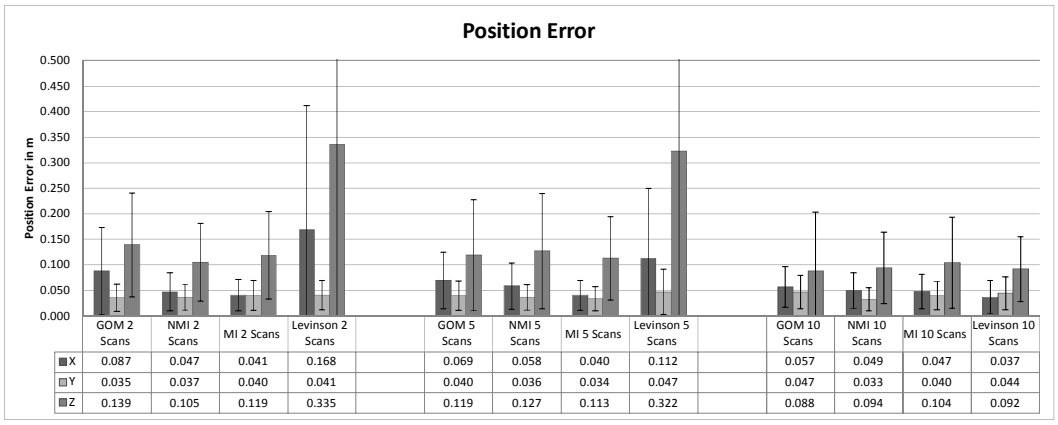
Fig. 5. Camera and velodyne scan being registered. Left, the velodyne scan. Centre, the Ladybug's centre camera image. Right the two sensor outputs overlaid.

All five cameras of the Ladybug were calibrated independently. An example of the process of registering one of the camera's outputs is shown in Figure 5. This calibration was performed 10 times for each camera using randomly selected scans each time. The error in each camera's relative position to each other camera in all trials was found and the average error shown in Table III.

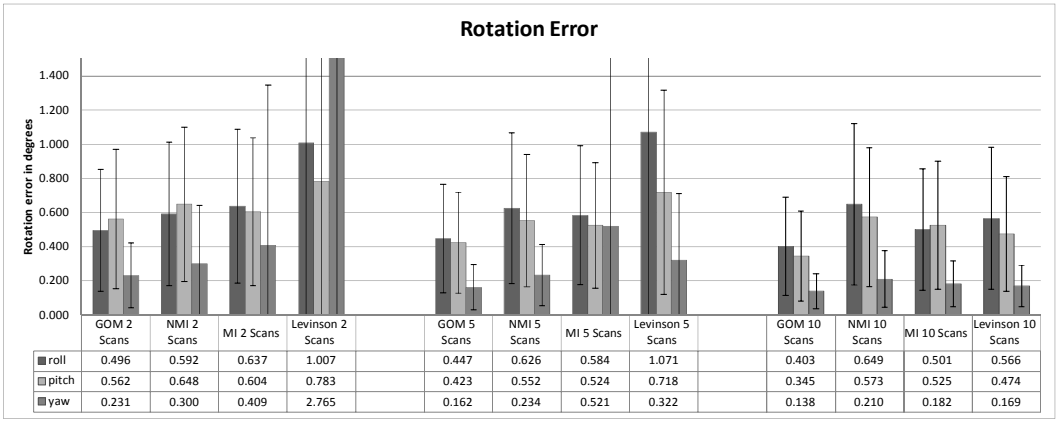
In these tests GOM, NMI and MI gave similar results. GOM tended to give the most accurate rotation estimates while MI gave the most accurate position estimates. For all three of these metrics, scan aggregation slightly improved the accuracy of angles and position. Levinson's presented the largest improvement in accuracy when more scans were aggregated, resulting in the largest error with 2 and 5 scans and giving similar results to the other methods with 10 scans.

In this experiment, any strong conclusion about which metric performed the best is difficult to draw as the difference between any two metrics for 10 aggregated scans is significantly less than the variance in their values. In almost all of the tests, the estimate of the cameras Z position was significantly worse than the X and Y estimates. This was expected as the metric can only be evaluated in the overlapping regions of the sensors fields of view. The Velodyne used has an extremely limited vertical resolution (64 points, one for each laser). Thus making the parallax error that indicates an error in the Z position difficult to observe. The narrow beam width of the Velodyne is also why the yaw shows the lowest error, as there are more overlapping points that can be used to evaluate this movement.

The actual error of a Ladybug-Velodyne system calibrated using all five cameras simultaneously would give a far more accurate solution than the results obtained here. There are several reasons for this. Individually the single camera systems have a narrow field of view. Therefore, a forward or backward translation of the camera is only shown through subtle parallax error in the position of objects in the scene. This issue is significantly reduced in the full system due to the cameras that give a perpendicular view that clearly shows this movement. In the single camera problem, movement parallel to the scene is difficult to distinguish from a rotation. This is also solved by the full system due to the very different effect a rotation and translation have on cameras facing in significantly different directions. Finally the full system also benefits from the increase in the amount of overlap



	GOM 2 Scans	NMI 2 Scans	MI 2 Scans	Levinson 2 Scans	GOM 5 Scans	NMI 5 Scans	MI 5 Scans	Levinson 5 Scans	GOM 10 Scans	NMI 10 Scans	MI 10 Scans	Levinson 10 Scans
X	0.087	0.047	0.041	0.168	0.069	0.058	0.040	0.112	0.057	0.049	0.047	0.037
Y	0.035	0.037	0.040	0.041	0.040	0.036	0.034	0.047	0.047	0.033	0.040	0.044
Z	0.139	0.105	0.119	0.335	0.119	0.127	0.113	0.322	0.088	0.094	0.104	0.092



	GOM 2 Scans	NMI 2 Scans	MI 2 Scans	Levinson 2 Scans	GOM 5 Scans	NMI 5 Scans	MI 5 Scans	Levinson 5 Scans	GOM 10 Scans	NMI 10 Scans	MI 10 Scans	Levinson 10 Scans
roll	0.496	0.592	0.637	1.007	0.447	0.626	0.584	1.071	0.403	0.649	0.501	0.566
pitch	0.562	0.648	0.604	0.783	0.423	0.552	0.524	0.718	0.345	0.573	0.525	0.474
yaw	0.231	0.300	0.409	2.765	0.162	0.234	0.521	0.322	0.138	0.210	0.182	0.169

TABLE III

Average error between two aligned Ladybug cameras. All distances are in metres and angles are in degrees.

between the sensors' fields of view.

V. Conclusion

We have presented a detailed evaluation of our *gradient orientation measure* (GOM). The measure can be used to align the output of two multi-modal sensors, and has been demonstrated on a variety of datasets and sensors. Three other existing methods were also implemented and their accuracy tested on the same datasets. On the datasets tested GOM successfully registered all datasets to a high degree of accuracy, showing the robustness of the method, for a large range of environments and sensor configurations. We also examined the level of accuracy required for an initial guess for a system's calibration to be optimised to the correct solution.

Acknowledgment

This work has been supported by the Rio Tinto Centre for Mine Automation and the Australian Centre for Field Robotics, University of Sydney.

References

[1] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *Medical Imaging, IEEE*, vol. 22, no. 8, pp. 986–1004, 2003.

[2] A. Mastin, J. Kepner, and J. Fisher III, "Automatic registration of LIDAR and optical images of urban scenes," *Computer Vision and Pattern Recognition*, pp. 2639–2646, 2009.

[3] J. Levinson and S. Thrun, "Automatic Calibration of Cameras and Lasers in Arbitrary Scenes," in *International Symposium on Experimental Robotics*, 2012.

[4] G. Pandey, J. R. Mcbride, S. Savarese, and R. M. Eustice, "Automatic Targetless Extrinsic Calibration of a 3D Lidar and Camera by Maximizing Mutual Information," *Twenty-Sixth AAAI Conference on Artificial Intelligence*, vol. 26, pp. 2053–2059, 2012.

[5] R. Wang, F. P. Ferrie, and J. Macfarlane, "Automatic registration of mobile LiDAR and spherical panoramas," *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–40, Jun. 2012.

[6] Z. Taylor, J. Nieto, and D. Johnson, "Automatic calibration of multi-modal sensor systems using a gradient orientation measure," *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1293–1300, Nov. 2013.

[7] J. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, 1965.

[8] S. Chen, "Another Particle Swarm Toolbox," 2009.

[9] G. Pandey, J. McBride, and R. Eustice, "Ford campus vision and lidar data set," in *The International Journal of Robotics Research*, 2011, pp. 1543–1552.

[10] Z. Taylor and J. Nieto, "A Mutual Information Approach to Automatic Calibration of Camera and Lidar in Natural Environments," in *the Australian Conference on Robotics and Automation (ACRA)*, 2012, pp. 3–5.