# Multi-modal sensor calibration using a gradient orientation measure

**Zachary Taylor**
University of Sydney, Australia
z.taylor@acfr.usyd.edu.au

**Juan Nieto**
University of Sydney, Australia
j.nieto@acfr.usyd.edu.au

**David Johnson**
University of Sydney, Australia
d.johnson@acfr.usyd.edu.au

## Abstract

This paper presents a new metric for calibrating multi-modal sensor data. The metric is based on the alignment of the orientation of image-intensity gradients formed from the two candidate sensors. It operates by finding the transformation that minimises the misalignment of these gradients. The metric can operate in a large range of applications working on both 2D and 3D sensor outputs. Unlike traditional calibration methods, it does not require markers or other registration aids to be placed in the scene. To demonstrate the effectiveness of the method we present results on a variety of camera-lidar and camera-camera registration problems. Comparisons with the state of the art methods are also presented. Our method is shown to give high quality registrations under all tested conditions.

## 1 Introduction

Robotic platforms typically carry a wide range of sensor modalities. Each of these sensors aim to provide the robot with different information about the surrounding environment. When the output from complementary sensors are combined, the robot is provided with a greater insight of the surroundings than any one sensor alone could provide. For example, two sensors whose outputs are commonly fused are cameras and lidar scanners to create 3D textured maps of the environment.

To integrate the information provided by two sensors, their relative location and orientation must be known. For precision applications this location cannot be found by simply measuring the sensors positions due to the uncertainty introduced in the measurement process and the uncertainty of the actual sensor location inside the casing. This calibration is far from trivial due to the very different modalities via which the two sensors may operate (Le and Ng, 2009). Because of these difficulties, on most mobile robots the sensors are manually calibrated. This is usually done using reflective markers, chequerboards or by painstakingly hand labelling large numbers of points. These methods are slow, labour intensive and often produce results with significant errors. Once this calibration has been completed it is assumed that it remains unchanged while the robot is operating. In practice however this is a poor assumption as the calibration is degraded due to the robot's motion, particularly for mobile robots working in rough environments. Therefore for these systems to be able to operate autonomously for extended periods of time a robust method that can automatically recalibrate the sensors without requiring the system to stop its current operations is required.

As a step towards achieving this goal this paper presents a new metric, the *gradient orientation measure* (GOM) that can be used to align the outputs of two sensors of different modalities. The metric can operate on both 3D-2D and 2D-2D registration problems. This makes the metric well suited to calibrating both camera-lidar and camera-camera systems, where the cameras may detect different frequencies of light. Unlike many

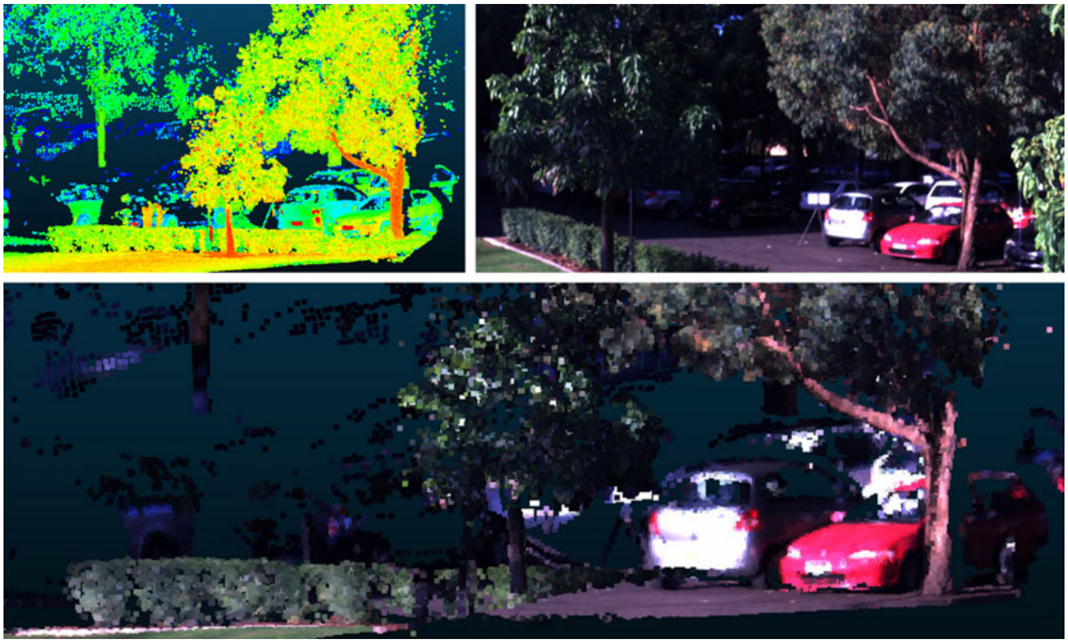


Figure 1: Camera and lidar scan being fused. Raw lidar data is shown in the top left, with the camera image shown in the top right. The textured map obtained with our approach is shown at the bottom.

approaches that calibrate lidar-camera systems, the metric is also able to calibrate from a single scan pair. To demonstrate the metric's potential and versatility we present results on three different datasets: (i) the calibration of a panospheric camera with a series of Velodyne scans, (ii) the calibration of a rotating panoramic camera with a single high resolution scan and (iii) the alignment of two hyper-spectral cameras. In each of these tests the proposed approach is compared to state of the art methods. An example of the results obtained with our system is shown in Figure 1.

Specifically, this paper presents the following contributions:

- A review of the state of the art in automatic sensor calibration.
- The introduction of a new metric for evaluating the alignment of two sensors of different modality.
- The evaluation of the proposed approach in three different datasets. The data are with different sensor modalities and in different types of environment.
- A comparison of our approach and three different techniques: Levinson and Thrun's method (Levinson and Thrun, 2012), Pandey *et al*'s method (Pandey et al., 2012) and normalised mutual information based method (Taylor and Nieto, 2012).
- The evaluation of different 3D features for use with the GOM method.

## 2 Related work

A large number of work has been done in multi-modal sensor calibration. To highlight the similarities and differences among them we will divide the related work into four sections. First we present a description of mutual information and then three application-based sections. We review methods that operate on two dense images. Then methods that match a single high resolution point cloud to an image. And finally methods that work on data gathered from a moving platform and optimise over a large number of data frames.

### 2.1 Mutual information

The most common multi-modal image matching technique used is known as mutual information (MI). MI is a measure of statistical dependency between two signals. It is widely used in medical image registration. A survey of MI-based techniques has been presented in (Pluim et al., 2003). MI is a core component of many multimodal registration techniques including two of the methods evaluated in this paper. Therefore we present next a brief description of the technique.

MI was first developed in information theory using the idea of Shannon entropy (Shannon, 1948). Shannon entropy is a measure of how much information is contained in a signal. Its discrete version is defined in Equation 1

$$H(X) = H(p_X) = \sum_{i=1}^{n} p_i log(\frac{1}{p_i}) \tag{1}$$

where $X$ is a discrete random variable with $n$ elements and the probability distribution $p_X = (p_1, ..., p_n)$.

If two distributions are independent then their joint distribution is equal to the sum of their individual distributions. As shown in Equation 2, MI uses this fact to give a measure of the signal's dependence by taking the difference between the independent and joint distributions of the entropy, where $H(M, N)$ is the joint entropy.

$$MI(M, N) = H(M) + H(N) - H(M, N) \tag{2}$$

When used for registration purposes MI can be influenced by the total amount of information contained in images causing it to favour images with less overlap (Studholme et al., 1999). This drawback is mitigated by using a *normalised mutual information metric* (NMI) defined in Equation 3.

$$NMI(M, N) = \frac{H(M) + H(N)}{H(M, N)} \tag{3}$$

In practice, for images, the required probabilities $p(M)$, $p(N)$ and $p(M, N)$ are typically estimated using a histogram of the distribution of intensity values.

### 2.2 Image-image systems

A vast number of methods have been proposed for solving the problem of multi-modal image matching and the related problem of multi-modal stereo correspondence. Many of these methods were first developed for the alignment of MRI and CT scans for use in medical imaging.

Bodensteiner *et al* successfully matched images using the common mono-modal image technique known as the scale invariant feature transform (SIFT) (Bodensteiner et al., 2011). In most multi-modal applications however it is found that the assumptions made about the directions and relative magnitudes of the gradients SIFT makes do not hold (Heinrich et al., 2012). To attempt to overcome this issue J. Chen developed a version of SIFT that was based on the absolute value of the gradient so that no distinction was made as to whether the gradient was increasing or decreasing (Chen and Tian, 2009).

Wachinger and Navab developed a method called entropy sum of squared differences (eSSD) that used the entropy of patches of the images for registration of T1 and T2 MRI scans. The method works by first creating images where each pixels intensity is equal to the entropy of the pixels in an n-by-n patch around it. Matching is then performed by taking the SSD of the two generated entropy images. In their tests they obtained results comparable to mutual information (Wachinger and Navab, 2010).

A method known as self similarity was initially developed by Shechtman and Irani (Shechtman and Irani, 2007) to identify an object in a scene from a rough sketch. It works by assuming that differently coloured areas in one image will be more likely to be coloured differently in the other modality. Several attempts have been made to use self similarity for multi-modal image matching, usually with slight changes to the implementation to increase performance. J. Huang et al used the measure to register LIDAR depth maps with aerial images of a city (Huang et al., 2011). Rather than using the approach in its standard form they used the descriptors as features to match points in each image. A similar approach was also applied in (Bodensteiner and Huebner, 2010), using SIFT points to register photos with IR and LIDAR images. Torabi *et al* used self similarity to perform multi-modal stereo correspondence between visual and IR images (Torabi and Bilodeau, 2011). Heinrich *et al* made use of an altered version they called the *modality independent neighbourhood descriptor* (MIND) (Heinrich et al., 2012) to register MRI with CT scans of the human brain. The main differences between MIND and self similarity are that the patches were a single pixel in size and no conversion to log-polar bins was made. The calculation of the variance was also simplified.

### 2.3 Single laser-image scan systems

A recently proposed method by H. Li *et al* makes use of edges and corners (Li et al., 2012). Their method works by constructing closed polygons from edges detected in both the lidar scan and images. Once the polygons have been extracted they are used as features and matched to align the sensors. The method was only intended for and thus tested using aerial photos of urban environments.

A. Mastin *et al* achieved registration of an aerial lidar scan by creating an image from it using a camera model (Mastin et al., 2009). 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.

A method for aligning ground based lidar scans of cliffs with hyperspectral photos of the same area has been developed by Nieto *et al* (Nieto et al., 2010). The method makes use of a pre-calibrated second camera that is rigidly attached to the lidar to give the lidar's scan points RGB colour. A camera model is then used to generate a colour image from this laser scan. The hyperspectral cameras image is matched to this generated image by using sift features to perform an affine transform on the image. The matching is then further refined using a local warping that utilizes the normalised cross correlation between patches.

While the above method for aligning photos of cliff faces with lidar scans worked well it had the drawback of requiring a second pre-calibrated camera. In our own earlier work we developed a method to perform the same task without this limitation (Taylor and Nieto, 2012). The method operated by creating an accurate camera model that emulated the hyperspectral camera and using it to project the lidar points onto the cameras images. Some of the lidar point clouds did not have usable intensity information and so a new intensity was assigned to the points based on an estimation of the direction of their normals. These lidar points were compared to the points in the image they were projected onto using normalised mutual information. It was assumed when this measure was maximised the camera model would have the same parameters as the actual camera used allowing it to relate each point in the image to its corresponding point in the lidars output point cloud.

For the alignment of fixed ground based scans in urban environments a large number of methods exist that exploit the detection of straight edges in a scene (Lee et al., 2002; Liu and Stamos, 2007). These straight lines are used to calculate the location of vanishing points in the image. While these methods work well in cities and with images of buildings, they are unable to correctly register natural environments due to the lack of strong straight edges.

From a more theoretical view on the calibration (Corsini et al., 2009) looked into different techniques for generating a synthetic image from a 3-D model so that MI would successfully register the image with a physical photo of the object. They used NEWUOA optimisation in their registration and looked at using the silhouette, normals, specular map, ambient occlusion and combinations of these to create an image that would robustly be registered with the real image. They found surface normals and a combination of normal and ambient occlusion to be the most effective.

### 2.4 Mobile systems

As the wide spread availability of 3D lidars capable of operating from a moving platform did not happen until relatively recently, the previous work in this field is rather limited. One of the first approaches that did not rely on markers was presented in (Levinson and Thrun, 2012). 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 blurred to increase the capture region of the optimizer. The average of all the edges 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 maximize the resulting metric.

Two very similar methods have been independently developed by Pandey *et al* and Wang *et al*. 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 mutual information 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 mutual information 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 the mutual information value from a large set of scans is required to allow the optimizers used to converge to the global maximum.

More recently an approach was developed by Napier *et al* for registering a push broom 2D lidar with a camera (Napier et al., 2013). To get an image from the 2D scanner its scans are first combined with an accurate navigation solution for the mobile system to generate a 3D scan. A 2D image is then produced from this 3D scan using a camera model. The two images have the magnitude of gradients present in them calculated and normalised over a small patch around them. The camera and lidar are assumed to be aligned when the sum of the differences in these gradient magnitude images are minimised. The metric also has an additional weighting that favours areas with higher resolution scans.

## 3 Multi-modal sensor calibration

### 3.1 System overview

Figure 2 illustrates the overall idea of our approach. The method can be divided into two main stages: feature computation and optimisation.

The feature computation stage converts the sensor data into a form that will allow for fast comparisons of different alignments during the optimisation stage. The initial step is the point intensity assignment, which is dependent on the dimensionality of the data produced by the sensors. For 2D data the average

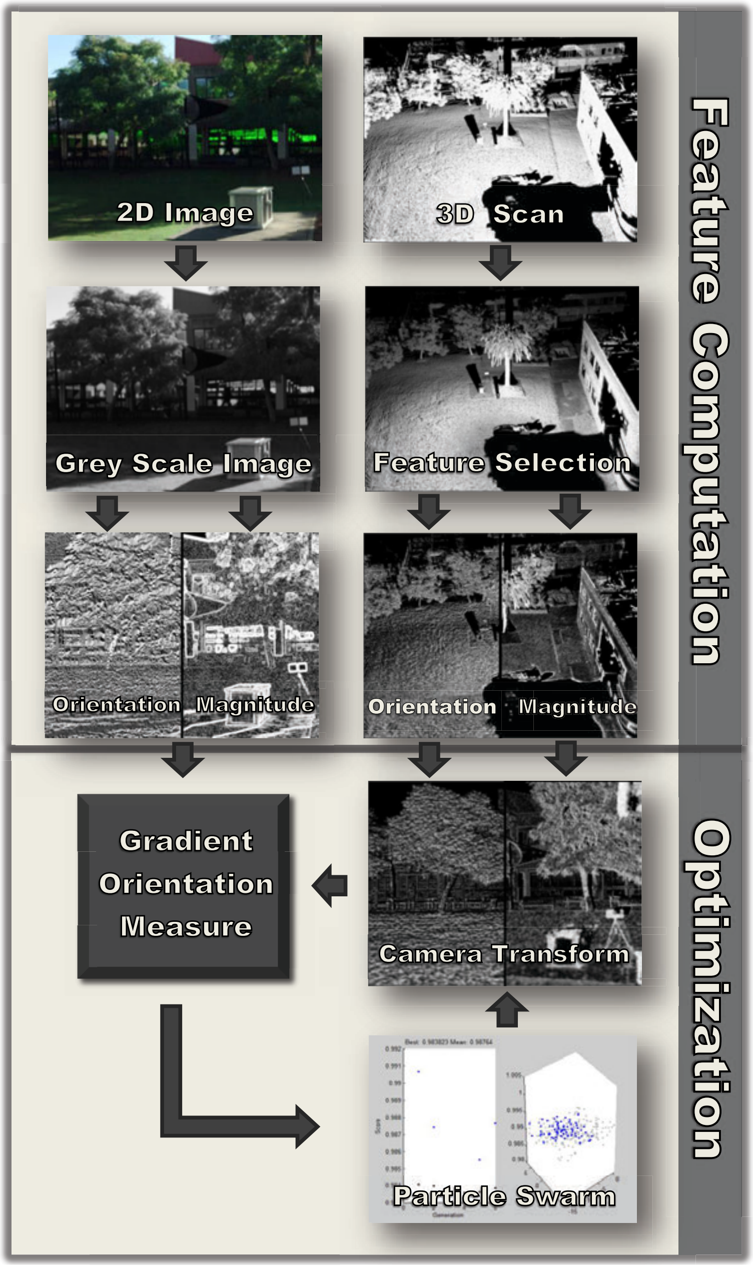


Figure 2: An example of the steps of our approach. This diagram shows the alignment of a camera image with a high resolution lidar scan coloured by its return intensity.

of the colour channels is used, however for 3D data the user selects one of several possible features usually depending on the exact sensor and situation (the features considered will be presented in section 3.2). The rest of the processing is independent of the dimensionality and proceeds as follows. Histogram equalisation is performed 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. 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 scan) and transforming the other scan (referred to as the moving scan). In our framework the base scan 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 moving scan onto the 2D base scan. Once this has been done the base scan is interpolated at the locations that the moving scan was projected onto to give the edge magnitudes and directions at these points.

GOM is used to compare the edge information from these points between the two scans and give a measure of how well aligned they are. This process is repeated for changing transformation parameters using particle swarm optimisation. The optimisation continues until all the particles converge and a global maximum for the GOM between the outputs is found. When multiple scans are taken the process performed is the same with the only difference that the GOM output is the average from comparing all of the scan pairs.[1]

### 3.2 3D features

For the chosen metric to correctly calibrate the system there has to be a strong relationship between the intensity of corresponding points. For the 3D sensors several possible features exist that can be used to set the points intensities. The features analysed in this work were the normals of the points, the return intensity and the distance of points from the sensors. Histogram equalisation is performed on all features to improve contrast.

**Normals of points**: To calculate an estimate of the normals a plane is approximated at the location of each point. This is done by first placing the points into a k-d tree. From this, the eight nearest neighbours to each point are found. The normal vector is calculated from the eigenvectors and eigenvalues of the covariance matrix $C$, given by equation 4 (Rusu, 2010).

$$C = \frac{1}{8}\sum_{i=1}^{8}(p_i - c)(p_i - c)^T \tag{4}$$

Where $p_i$ is the i-th nearest neighbour location and c is the location of the point. The smallest eigenvalue of $C$'s corresponding eigenvector is the best estimate of the normal vector to the plane. Once the normals have been calculated the three values that make up the normals are converted into a single intensity value by calculating the difference in angle between the normal vector and a line between the point and the origin of the scan. While other methods based on the angle of the points can be used, this method was empirically found to give good results.

**Return intensity**: Lidars provide a measure of the return strength of the laser from each point. This usually gives a strong relationship between the intensity of matching points as both laser reflectance and the camera pixel intensity primarily rely on the reflectance of the target material. While most radar and lidar systems will output the intensity reading it cannot be used in all situations. As the intensity of return is dependent on the distance of the object from the sensor, when multiple scans from different location are combined the readings cannot be used. A second issue with this method occurs when using Velodyne scans.

[1] All the code used for our method as well as additional results and documentation is publicly available online at http://zacharytaylor.github.io/Multimodal-Calib/

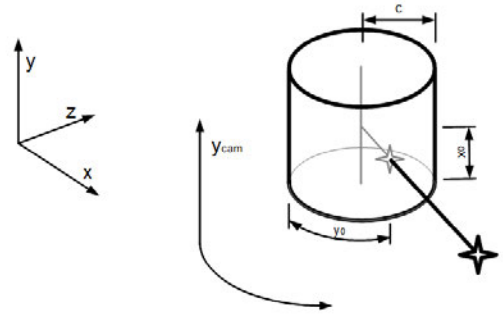


Figure 3: Cylinder model used to represent a panoramic camera

In a velodyne the scanning of the environment is done by 64 separate lasers. Each of these lasers has slightly different characteristics and can give substantially different intensity of returns for the same object.

**Distance of points:** The distance from the scanner to a point is a simple fast and often effective way of generating intensity values for 3D points. The feature works best when used in fairly cluttered scenes with a large number of objects at different distances from the camera such as on a busy street or in a garage. In open environments such as fields or highways however the method generally fails.

### 3.3 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.

#### 3.3.1 Camera models

To convert the data from a list of 3D points to a 2D image that can be compared to a photo, the points are first passed into a transformation matrix that aligns the sensor's axis. After this has been performed, one of two basic camera models is used. For most sensors a pin-hole camera model is used as defined in Equation 5. For some of our datasets the images were obtained from a panoramic camera. Panoramic cameras operate slightly differently to regular cameras as the image of the world is built up by rotating a single vertical line array. To account for this, a camera model that projects the points onto a cylinder must be used. A rough depiction of this is shown in Figure 3. This model projects the points using Equation 6 (Schneider and Maas, 2003),

$$x_{cam} = x_0 - \frac{cx}{z} + \Delta x_{cam}, \quad y_{cam} = y_0 - \frac{cy}{z} + \Delta y_{cam} \tag{5}$$

$$x_{cam} = x_0 - c\arctan(\frac{-y}{x}) + \Delta x_{cam}, \quad y_{cam} = y_0 - \frac{cz}{\sqrt{x^2+y^2}} + \Delta y_{cam} \tag{6}$$

where

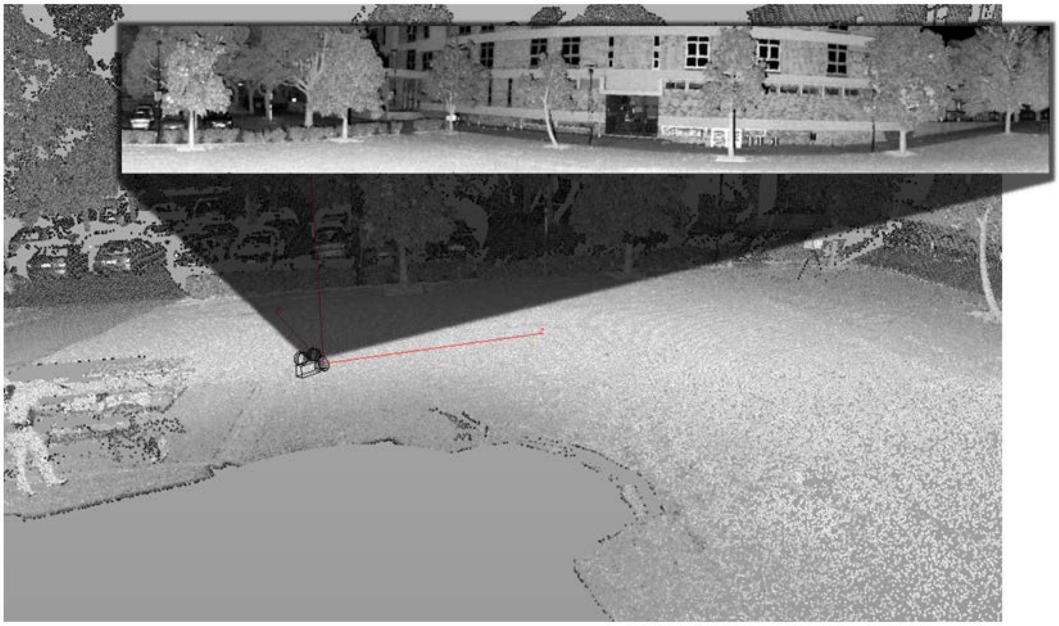


Figure 4: Image generated by placing the camera model in a lidar point cloud

$\mathbf{x}_{cam}$ , $\mathbf{y}_{cam}$ are the x and y position of the point in the image.
**x, y, z** are the coordinates of points in the environment.
**c** is the principle distance of the model.
$\mathbf{x}_0$ , $\mathbf{y}_0$ are the location of the principle point in the image.
$\Delta\mathbf{x}$ , $\Delta\mathbf{y}$ are the correction terms used to account for several imperfections in the camera.

A depiction of how the camera model operates on a point cloud can be seen in Figure 4.

### 3.3.2 Affine transform

The registration of two camera images depends on the distance of objects from the cameras. As the camera does not provide a depth measurement this means that if two cameras were to be perfectly aligned the registration parameters used would need to be recalculated every frame, as is done in the production of stereo images. However as the aim of this process is to produce simple calibration parameters that can be used to fuse modalities an affine transform is used. While not perfect, for two cameras with only small differences in location and orientation an affine transform can be used to give a high quality image registration.

## 3.4 Projecting 3D points to 2D images

Several issues arise when attempting to create an image from the point cloud produced by a 3D sensor. The sparse nature of the scans (especially those obtained from mobile platforms) mean that the majority of the pixels have no associated intensity value. Those pixels that are occupied often have more than one associated point creating a many-to-one mapping resulting in a significant loss of information. Aliasing issues also occur from forcing the points onto the discrete grid that makes up an image. These issues significantly degrade the quality of the alignments, especially for the GOM method as the aliasing issues seriously impacted the calculated edge directions required. A typical image produced from velodyne data is shown in Figure 5. To overcome these issues a range of different post processing options and interpolation or blurring techniques were attempted. However it was found that most of these techniques would destroy sharp edges and do little

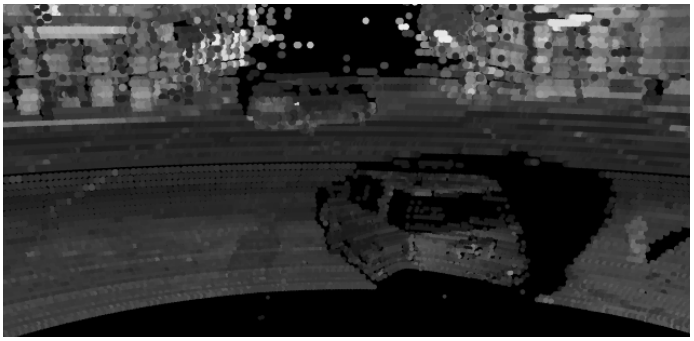


Figure 5: Image generated from velodyne lidar data. The sparse points lead to a poor quality image from which gradients cannot be accurately calculated.

to improve results.

To prevent these issues the generation of a traditional image from the 3D data is only done for visualisation purposes. Instead the points are kept in a list and when they are projected using the camera model their position is not discretised. To get the matching points from the base sensor's image, cubic spline interpolation is performed at the coordinates given by the point list.

### 3.5 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. We want the gradient of the points in the 3D data to be from the perspective of the camera so that the orientations for both sensors will be aligned once the camera models have been used to transform the data. The simplest way to calculate the gradients for the 3D data is to first use the camera model to generate an image and use a standard Sobel operator on it. Unfortunately this method gives poor results, due to the problems outlined in section 3.4

Instead the gradient of the 3D points are calculated as follows. The points are first projected onto a sphere that is centred at the estimated location of the camera using Equations 7 and 8.

$$x_{sphere} = arccos(\frac{z}{\sqrt{x^2+y^2+z^2}}) \tag{7}$$

$$y_{sphere} = arctan(\frac{y}{x}) \tag{8}$$

A sphere is used rather than the plane used in image generation, as with a plane, points in front of and behind the camera can be projected onto the same location. Each point on the sphere has the 8 nearest neighbours to it calculated before the gradient is calculated using Algorithm 1.

As the gradient is dependent on the location of the camera it requires re-estimation every time the camera's extrinsic parameters are changed. However as this process is fairly computational expensive for the purpose of gradient calculation in our process it is assumed that $parameters_{initial} \approx parameters_{final}$. This assumption allows the gradients to be pre-calculated and gives a massive reduction in the computational cost. In our experiments this assumption did not appear to negatively impact the accuracy of the method.

```
**Let**
$p_x, p_y, p_v$ be the current points x-position, y-position and intensity-value, respectively
$n_x, n_y, n_v$ be the neighbouring points x-position, y-position and intensity-value, respectively
$g_x, g_y, g_{mag}, g_{or}$ be the gradient in the x-direction, gradient in the y-direction, gradient-magnitude and orientation, respectively

$g_x = 0;$
$g_y = 0;$
$g_{mag} = 0;$
**for** *neighbouring point n* **do**
    $x = p_x - n_x;$
    $y = p_y - n_y;$
    $v = p_v - n_v;$
    $g_x = g_x + \frac{vx}{8};$
    $g_x = g_y + \frac{vy}{8};$
    $g_{mag} = g_{mag} + |\frac{v}{8}|;$
**end**
$g_{or} = arctan2(g_y, g_x)$
```

**Algorithm 1:** gradient calculation for 3D sensors

### 3.6 The Gradient Orientation Measure

The formation of a measure of alignment between two multi-modal sources is a challenging problem. Features in one source can be weak or missing from 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 chance there will be a large change in intensity in the other modality. This correlation exists as these large changes in intensity usually occur because of a difference in the material or objects being detected; changes which generally affect most sensor modalities.

GOM exploits these differences to give a measure of the alignment. GOM was inspired by a measure proposed by Josien P. W. *et al* in (Pluim et al., 2000) for use in medical imaging registration. The presented measure however has significant differences as Josien P. W. *et al's* method is un-normalised, uses a different calculation of the gradients strength and is combined with mutual information.

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 their direction $\alpha$ is by using Equation 9.

$$\alpha_j = cos(2(or_{1,j} - or_{2,j})) + 1 \tag{9}$$

Where $or_{i,j}$ is the orientation of the gradient in image i at point j. The difference in angle is multiplied by two 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. This means that for two aligned images the two corresponding gradients may be out of phase by 180 degrees.

Sharper gradients represent features that are more likely to be preserved between images. The stronger gradients also mean that the direction of the gradient calculated will be less susceptible to noise and thus more accurate. This means that these points should be given an increased weight. This is accomplished by multiplying $\alpha$ by a factor $\mu$. Where we define $\mu$ as the product of the two gradient magnitudes at that point as shown in Equation 10.

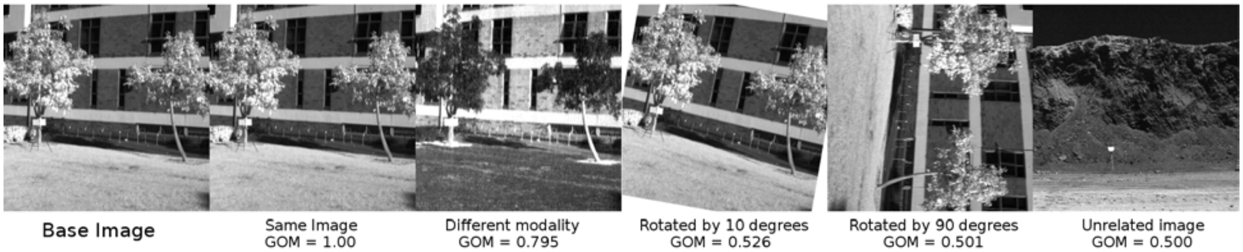


Figure 6: GOM values when the base image shown on the left is compared with a range of other images.

$$\mu_j = mag_{1,j} * mag_{2,j} \tag{10}$$

Where $mag_{i,j}$ is the magnitude of the gradient in image i at point j. Summing the value of all 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 the measure is normalised by dividing it by the sum of all of the gradient magnitudes, $\mu$. This gives the final measure which is shown in Equation 11.

$$GOM = \frac{\sum_{j=1}^{n} \mu_j \alpha_j}{2 \sum_{j=1}^{n} \mu_j} \tag{11}$$

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. Something of note is that if the two images were completely uncorrelated we would expect the measure to give a value of 0.5. This means that if two images have a GOM value of less than 0.5 the score is worse than random and it is a fairly safe assumption that the system is in need of calibration. Some typical GOM values for a range of images is shown in Figure 6.

# 4 Optimisation

Optimisation of GOM and other similar multimodal measures such as NMI or the Levinson method is a very challenging problem. There are several reasons for this difficulty. Generally the evaluation of the measure is computationally expensive due to the large number of points that are being compared. The extrinsic calibration of a camera is a 6 degree of freedom problem with a large range in which the correct solution can lie. Because of this expense of evaluation and large search space any attempt at an exhaustive search method is computational infeasible.

Depending on the sensors and the number of frames available the problem can be highly non-convex with large numbers of distinct local maximums, an example of a typical search space plotted over two dimensions is shown in Figure 7. These local maximums are caused by a large number of factors, in many cases they are caused by several strong edges of objects happening to align with incorrect objects, this is often due to the repetitive nature of the construction of man made structures. For example when experimenting with optimizers it was commonly found that one of our scans of a building would have the first floor in the lidar scan aligned with the ground floor in the image due to the similar structure. Averaging GOM values over a large number of scans reduces the impact of these chance overlap of strong edges to the point where gradient accent methods may be used if a fairly accurate initial guess is known, however in our experience these optimizers are of limited use.

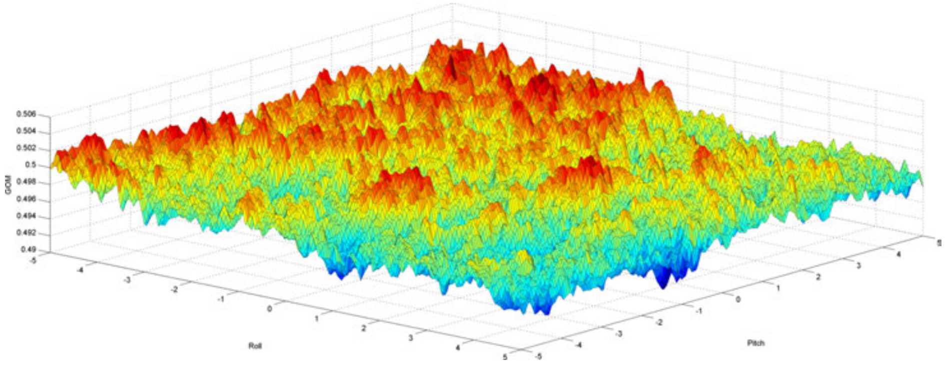


Figure 7: GOM values plotted for the roll and pitch of a typical lidar-photo alignment.

To find the optimum value in a reasonable range of possible solutions we found through trial particle swarm was the most effective optimizer for solving our problem reliably. Because of this we used it for finding the optimal values for the GOM, NMI and Levinsion method in our tests.

# 5 Experimental Results

## 5.1 Implementation

The code for testing the methods outlined was written in Matlab with the exception of the three most computationally expensive sections. These sections, the transformation of the point cloud, interpolation of images and the evaluation metrics were coded in CUDA to reduce computation time[2]. Running on a desktop with an i7-875k CPU and a GTX470 GPU the transformation, interpolation and GOM evaluation of a Velodyne scan containing 80,000 points with a camera image requires roughly 4ms. The process scales linearly with the number of points in the scans and the number of images; it is independent of the resolution of the image. To register 20 Velodyne scans with 100 Ladybug images requires approximately 3 to 5 hours for a particle swarm optimisation with 200 particles.

## 5.2 Parameters Optimisation

To optimise the parameters, our implementations of the NMI, GOM and Levinson methods require us to define a search space. We construct this search space around an initial guess. The initial guess we use is either the ground truth value (when available) or a manually calibrated solution. We then added a random offset to it. The random offset is set to be uniformly distributed and up to half the size of the search space. 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.

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 non-deterministic optimisers were run 10 times with the mean and standard deviation from these runs reported for each dataset.

[2] All the code for the implementation as well as documentation is publicly available at http://zacharytaylor.github.io/Multimodal-Calib/.

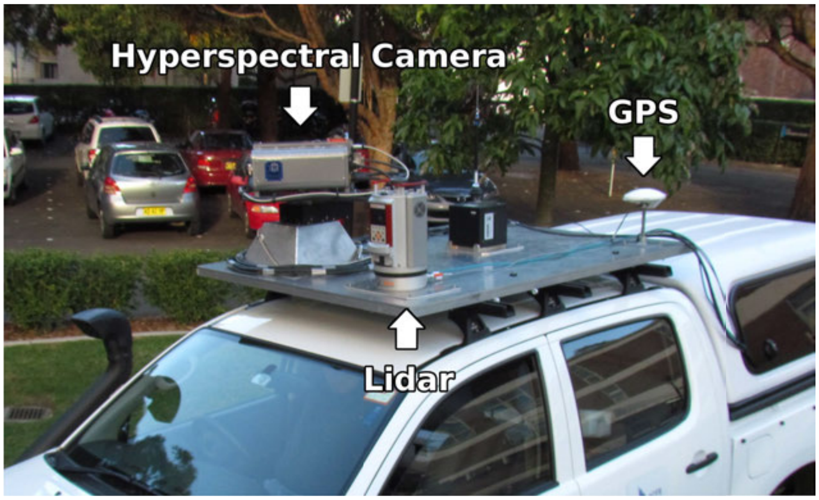


Figure 8: Setup used to collect ACFR data

### 5.3 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 from the grass courtyard next to it. The setup is shown in Figure 8. The scanner output gave the location of each point as its latitude, longitude and altitude. The focal length of the hyper-spectral camera was adjusted between each scan.

This dataset required the estimation of an intrinsic parameter of the camera, its focal length in addition to its extrinsic calibration. As the code used to test G. Pandey *et al*'s method did not account for this parameter when it was run the focal length from the most accurate of the other tested methods was provided to it. The code was also modified slightly to allow it to be used with a panoramic camera.

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 cameras principal distance was within 40 pixels of correct (for this camera principal distance $\approx 780$) and the x, y and z coordinates were within one metre of correct.

#### 5.3.1 Results

No accurate ground truth is available for the ACFR dataset. To overcome this issue and allow an evaluation of the accuracy of the method, 10 points in each scan-image pair were matched by hand. An example of this is shown in Figure 9. An evaluation as to the accuracy of the method was made by measuring the distance in pixels between these points on the generated images. The results of this are shown in Table 1.

For this dataset GOM significantly improved upon the initial guess for all four of the tested scans. The NMI solution however gave poor results only improving on the initial point in two cases and giving significantly worse results in one case. G. Pandey *et al* method also performed poorly on this dataset usually giving results not significantly different than the initial conditions. This was probably due to the optimiser used quickly finding and becoming trapped in a local maximum near where it started. This would have occurred as the method was designed primarily for use on groups of scans simultaneously (as in the Ford dataset) relying on the combination of the scans to smooth out these local maximums. Levinson's method was not used on this dataset as it requires multiple images to operate due to a step that takes the difference between an image and the average image.

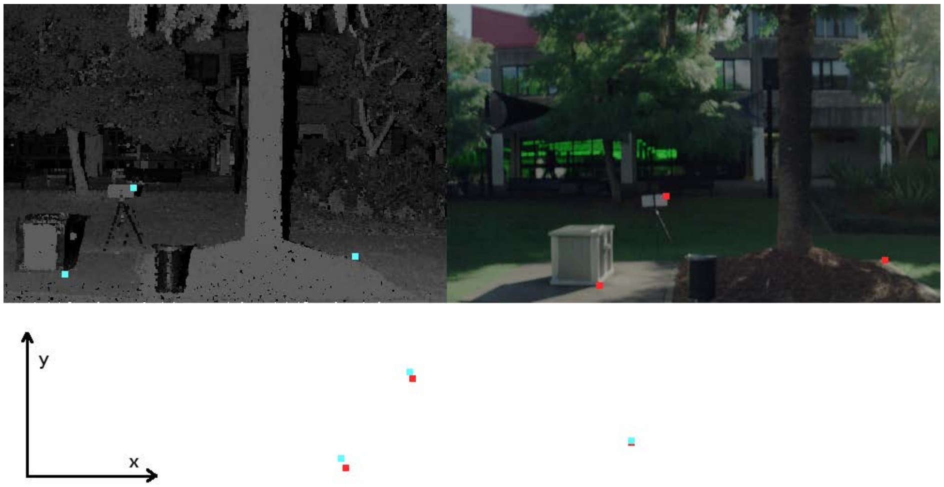


Figure 9: Hand labelled points for a section of image-scan pair 1 aligned by GOM. The top left image shows the lidar scan with three of the labeled points highlighted in blue. The top right image shows the camera image with the three corresponding points highlighted in red. The bottom image shows the two sets of points overlaid so that the registration accuracy can be judged.

| Scan | Initial | GOM | NMI | G. Pandey *et al* |
|---|---|---|---|---|
| 1 | 37.4 | 10.9 | 117.3 | 46.4 |
| 2 | 21.9 | 5.9 | 24.2 | 26.7 |
| 3 | 12.5 | 6.7 | 12.1 | 13.9 |
| 4 | 22.9 | 3.8 | 13.4 | 9.6 |
| **Average** | **23.7** | **6.9** | **39.9** | **24.1** |

Table 1: Accuracy comparison of different methods on ACFR dataset. All distances in pixels

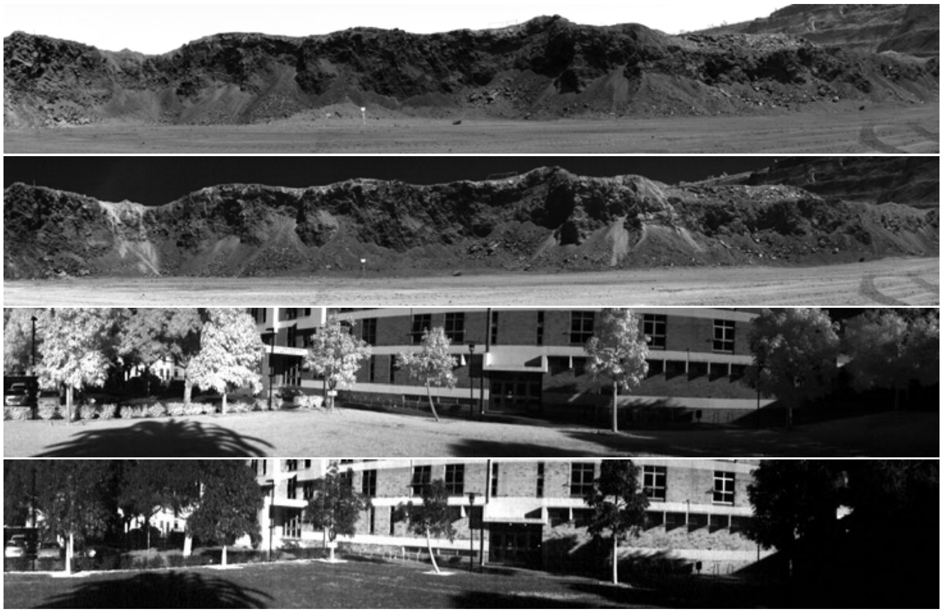


Figure 10: Images captured by hyper-spectral camera. From top to bottom the image intensity corresponds to bands: 420nm, 950nm, 1010nm and 2005nm

### 5.4 Dataset II

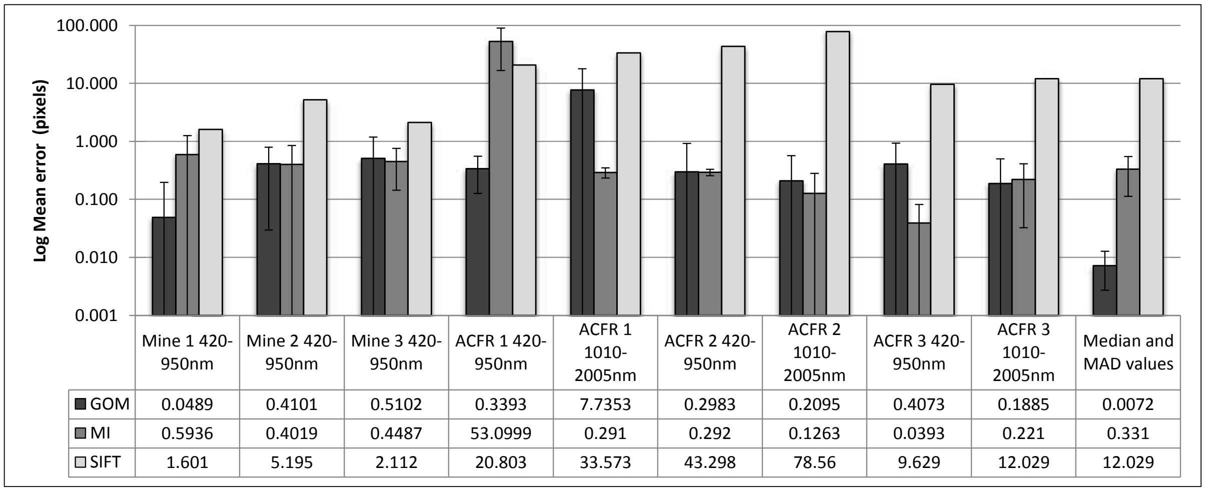
To test the methods ability to register different modality camera images such as IR-RGB camera alignment, two scenes were scanned with a pair of hyper-spectral cameras. Hyper-spectral camera captures images from a large number of light wavelengths at the same time. This made them ideal for testing the methods accuracy as the different modality images are near perfectly aligned to start with as all the different wavelength images are captured at the same time onto the same CCD. This removes any difference in the camera intrinsics or extrinsics and means that a perfect ground truth exists and it is easy to quantify any error a registration method has.

For each of the two hyper-spectral cameras available (VNIR and SWIR), bands near the upper and lower limits of the cameras' 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. For the first camera the bands selected were at 420 nm (violet light) and 950 nm (near ir), while for the second camera 1010 nm (near IR) and 2005 nm (near IR) were chosen. Both cameras were used to take a series of three photos of the ACFR building to test the approach on. Camera one was also used to capture three images of cliff faces at a mine site. An example of the images taken is shown in figure 10.

The search space for the particle swarm optimiser was setup assuming the x and y translation were within 30 pixels of the actual image, the rotation was within 10 degrees of the actual image, the x and y scale were within 20% of the actual image and the x and y shear were within 5% of the actual image.

#### 5.4.1 Results

The results of the image alignment are summarised in Table 2. As well as the GOM and MI methods that have been applied to all the datasets SIFT features were also used. SIFT was used to match feature points and combined with RANSAC to give the final transform used to give the results. 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 2. The images taken at



| | Mine 1 420-950nm | Mine 2 420-950nm | Mine 3 420-950nm | ACFR 1 420-950nm | ACFR 1 1010-2005nm | ACFR 2 420-950nm | ACFR 2 1010-2005nm | ACFR 3 420-950nm | ACFR 3 1010-2005nm | Median and MAD values |
|---|---|---|---|---|---|---|---|---|---|---|
| ■ GOM | 0.0489 | 0.4101 | 0.5102 | 0.3393 | 7.7353 | 0.2983 | 0.2095 | 0.4073 | 0.1885 | 0.0072 |
| ▨ MI | 0.5936 | 0.4019 | 0.4487 | 53.0999 | 0.291 | 0.292 | 0.1263 | 0.0393 | 0.221 | 0.331 |
| □ SIFT | 1.601 | 5.195 | 2.112 | 20.803 | 33.573 | 43.298 | 78.56 | 9.629 | 12.029 | 12.029 |

Table 2: Error and standard deviation of different registration methods performed on the hyper-spectral cameras' images. Error is given as the mean per-pixel-error in position. Note that the charts axis uses a log scale.

the ACFR were 320 by 2010 pixels in size. The width of the images taken at the mine varied slightly but were generally around 1600 by 8000 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 so 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 and GOM gave similar performance on these datasets both achieving sub-pixel accuracy in most cases. MI typically gave slightly more accurate results and there was significantly less variation in the results from each of the runs. Both MI and GOM had one instance of mis-registration on the ACFR 1 scan; in both of these cases most of the runs found an accurate solution, however one or two converged to a result with significant error. ACFR 1 was the most challenging scan used as large shadows cast over most of the image completely obscured many of the features in the image.

When summarizing the overall error of the methods on this dataset we chose to report the median and MAD (median absolute deviation) absolute error, rather then the mean and standard deviation. This decision was made as generally the methods converged to accurate solutions, however on the rare occasion when the methods failed, they would often converge to solutions with hundreds of pixels error. Because of this we believe the median shows a more accurate representation of the typical performance of the methods. Under this measure it can be seen that the GOM method gives highly accurate results with a median error of less then a hundredth of a pixel for this dataset.

### 5.5 Dataset III

The Ford campus vision and lidar dataset is a dataset provided by G. Pandey *et al* (Pandey et al., 2011). The test rig was a Ford F-250 pick-up truck that had a Ladybug panaspheric camera and Velodyne lidar

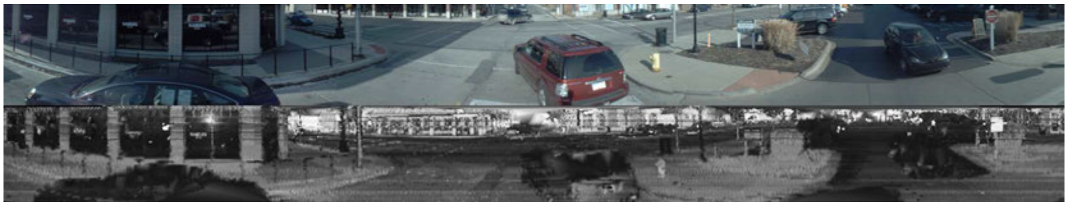


Figure 11: Overlapping region of camera image (top) and lidar scan (bottom) for a typical scan-image pair in the Ford Dataset. The lidar image is colour by intensity of laser return

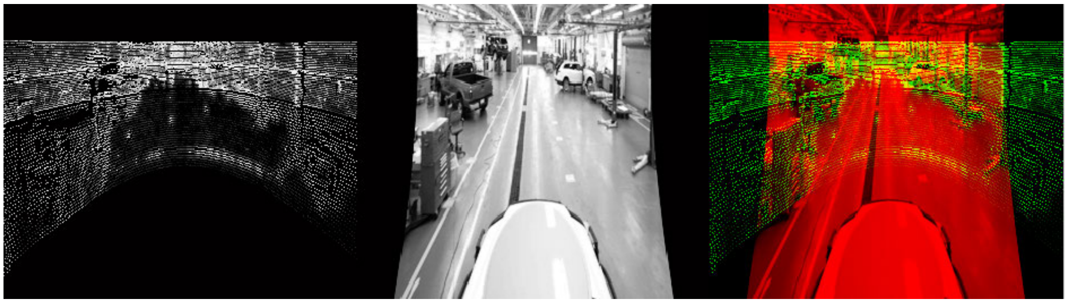


Figure 12: Camera and velodyne scan being registered. Left, the velodyne scan. Centre, the ladybugs centre camera image. Right the two sensor outputs overlaid.

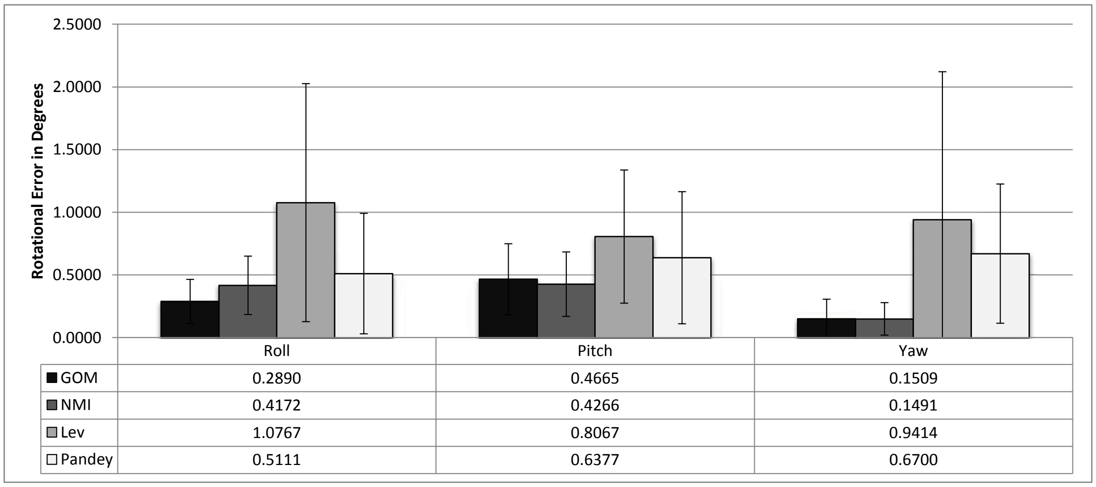
mounted on top of it. The dataset contains scans obtained driving around downtown Dearborn, Michigan USA. The testing was performed on this dataset as it offers a variety of environments and the Velodyne scanner used in this test has been calibrated to account for the different return characteristics of each laser. An example of the data is shown in Figure 11. It is also the dataset G. Pandey *et al*'s method was developed on making it the ideal dataset on which to test our method against. 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 G. Pandey *et al*'s method. The search space used for the particle swarm optimiser assumed the roll, pitch and yaw of the camera were within 3 degrees of correct and the x, y and z coordinates were within 0.5 metres of correct

#### 5.5.1 Results

The Ford dataset does not have a ground truth to compare the results of the calibration against. However a measure of the 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.

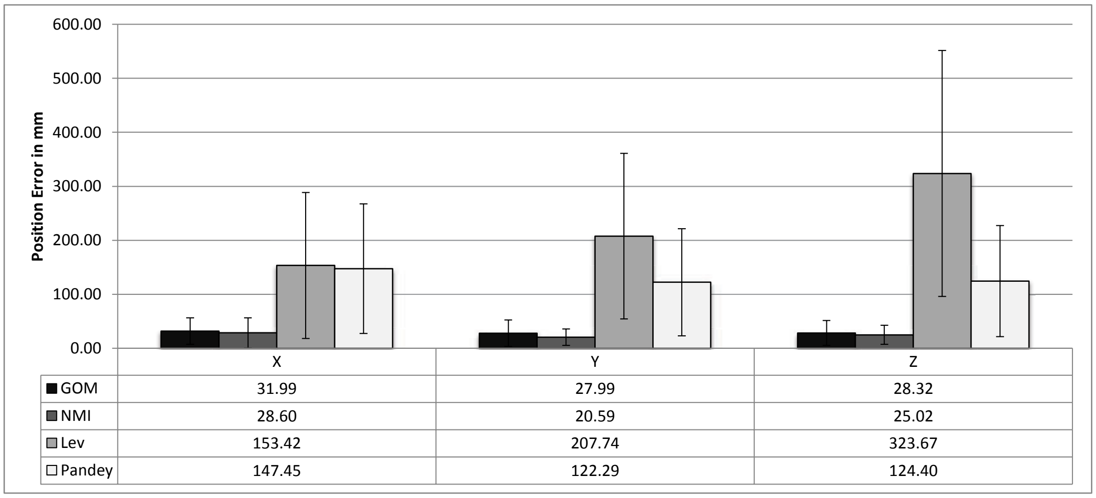
All five cameras of the Ladybug were calibrated independently. An example of the process of registering one of the cameras' outputs is shown in Figure 12. This calibration was performed 10 times for each camera. The error in each camera's relative position to each other camera in all trials was found (1000 possible combinations) and the average error shown in Table 3 and 4.

In these tests GOM and NMI both performed well giving very similar accurate results. GOM gave slightly more accurate angles while NMI gave more accurate distances, however the difference between the methods is small making it difficult to determine which method performed better. *Pandey et al* method gave okay performance, however it performed significantly worse than the NMI method despite the metrics used being very similar. This implies the error in the initial location must have been too large for the optimiser, placing



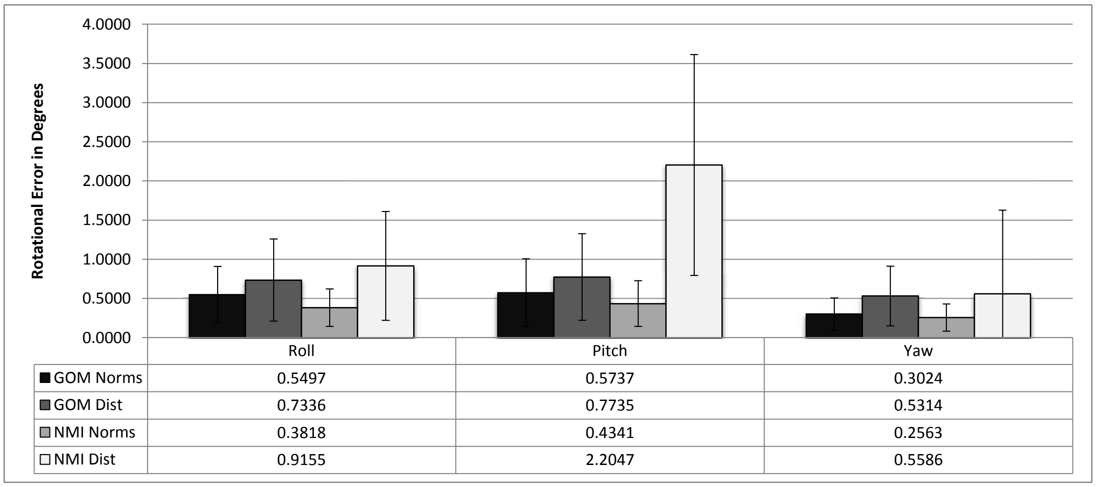
| | Roll | Pitch | Yaw |
|---|---|---|---|
| ■ GOM | 0.2890 | 0.4665 | 0.1509 |
| ■ NMI | 0.4172 | 0.4266 | 0.1491 |
| ■ Lev | 1.0767 | 0.8067 | 0.9414 |
| □ Pandey | 0.5111 | 0.6377 | 0.6700 |

Table 3: Average rotational error between two aligned ladybug cameras. All angles in degrees



| | X | Y | Z |
|---|---|---|---|
| ■ GOM | 31.99 | 27.99 | 28.32 |
| ■ NMI | 28.60 | 20.59 | 25.02 |
| ■ Lev | 153.42 | 207.74 | 323.67 |
| □ Pandey | 147.45 | 122.29 | 124.40 |

Table 4: Average translational error between two aligned ladybug cameras. All distances in millimetres



| | Roll | Pitch | Yaw |
|---|---|---|---|
| ■ GOM Norms | 0.5497 | 0.5737 | 0.3024 |
| ■ GOM Dist | 0.7336 | 0.7735 | 0.5314 |
| ■ NMI Norms | 0.3818 | 0.4341 | 0.2563 |
| □ NMI Dist | 0.9155 | 2.2047 | 0.5586 |

Table 5: Average rotational error between two aligned Ladybug cameras for different 3D features. All angles in degrees

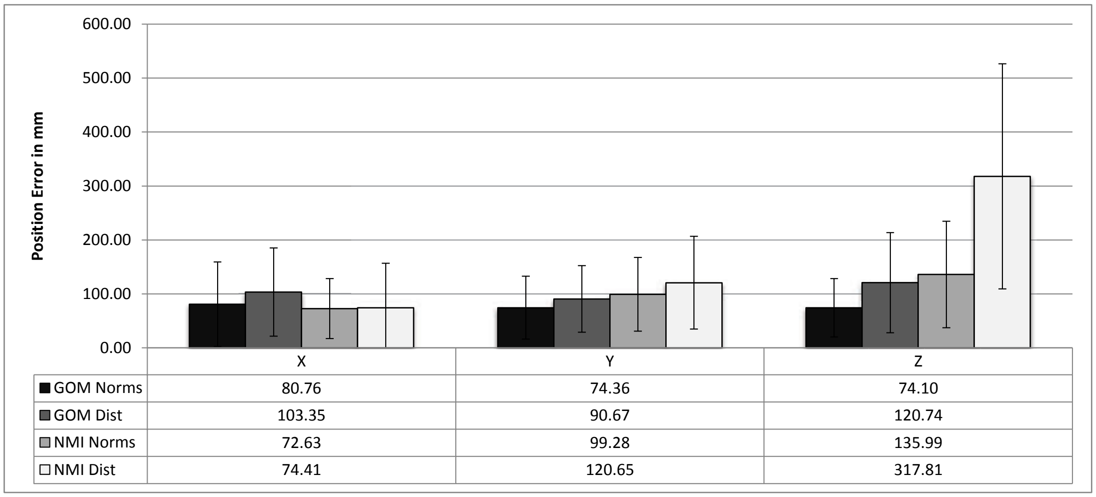
it outside the global maximums domain of attraction. In this test the Levinson method gave unusual results; it offered fairly accurate values for its prediction of rotations, but its performance concerning the location of the camera was worse than random for both the Y and Z position. This may have been caused as the metric's un-normalised nature meant it had a bias to placing the camera as far back as possible to maximise the overlap between the camera's field of view and the lidar's scan.

While all of the methods tested have a somewhat large error, the actual error of a Ladybug-Velodyne system calibrated using all five cameras simultaneously would give a far more accurate solution. 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 between the sensors' fields of view.

### 5.6 Feature comparison

With many sensors the rich return intensity information provided by these lidar would not be available. This may be due to the fusing of multiple scans, the sensor having a low number of intensity bits (many lidar only have 3 bit return intensity) or the sensor returns tuned to detect corner reflectors giving an almost zero intensity for all other points. In these situations the alternative features suggested in section 3.2 must be used. To test their accuracy the Ford dataset was evaluated using these different features, with all other parameters kept the same for both NMI and GOM.

The results of these tests are presented in Tables 5 and 6. Under these conditions the normals feature was found to give the most accurate results for both measures. Using normals the NMI method gave the most accurate location and the GOM method gave the most accurate rotation. While all of these results are worse than those obtained using the return intensity, the results obtained using the normals are accurate enough to provide a viable option in circumstances where the intensity information is not available.



| | X | Y | Z |
|---|---|---|---|
| ■ GOM Norms | 80.76 | 74.36 | 74.10 |
| ▩ GOM Dist | 103.35 | 90.67 | 120.74 |
| ▨ NMI Norms | 72.63 | 99.28 | 135.99 |
| □ NMI Dist | 74.41 | 120.65 | 317.81 |

Table 6: Average translational error between two aligned Ladybug cameras for different 3D features. All distances in millimetres

### 5.7 Methods conclusions

From the experiments performed and using all of the different methods we have found there are situations where each of the method's strengths make it the most appropriate to use. When given a sufficiently accurate initial condition Pandey *et als* method will converge to the correct result and operate an order of magnitude quicker than the other methods due to its local optimiser. When large amounts of data are presented and there is a strong correlation between intensities in the two sensors' outputs the NMI method gives good results. The Levinson method gives some of the most accurate results in situations where the return intensity is not available. In all tested situations the GOM method was found to give a consistently accurate estimate of the correct values.

## 6 Conclusion

We have introduced a new technique for multi-modal registration, the 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 dataset. On the datasets tested GOM successfully registered all datasets to a high degree of accuracy, showing the robustness of the method in a large range of environments and sensor configurations. Finally we explored different features that could be used in combination with GOM or the mutual information technique tested to register the outputs.

## 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

Bodensteiner, C., Hubner, W., and Jungling, K. (2011). Monocular Camera Trajectory Optimization using LiDAR Data. *Computer Vision*, pages 2018–2025.

Bodensteiner, C. and Huebner, W. (2010). Local multi-modal image matching based on self-similarity. In *Image Processing (ICIP)*, pages 937–940.

Chen, J. and Tian, J. (2009). Real-time multi-modal rigid registration based on a novel symmetric-SIFT descriptor. *Progress in Natural Science*, 19(5):643–651.

Corsini, M., Dellepiane, M., Ponchio, F., and Scopigno, R. (2009). Image to Geometry Registration: a Mutual Information Method exploiting Illumination-related Geometric Properties. *Computer Graphics Forum*, 28(7):1755–1764.

Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, S. M., and Schnabel, J. a. (2012). MIND: Modality Independent Neighbourhood Descriptor for Multi-Modal Deformable Registration. In *Medical Image Analysis*, pages 1423–1435. Elsevier B.V.

Huang, J., You, S., and Zhao, J. (2011). Multimodal image matching using self similarity. In *Applied Imagery Pattern Recognition (AIPR)*, pages 1–6.

Le, Q. V. and Ng, A. Y. (2009). Joint calibration of multiple sensors. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3651–3658.

Lee, S., Jung, S., and Nevatia, R. (2002). Automatic integration of facade textures into 3D building models with a projective geometry based line clustering. In *Computer Graphics Forum*, volume 21, pages 511–519.

Levinson, J. and Thrun, S. (2012). Automatic Calibration of Cameras and Lasers in Arbitrary Scenes. In *International Symposium on Experimental Robotics*.

Li, H., Zhong, C., and Huang, X. (2012). Reliable registration of LiDAR data and aerial images without orientation parameters. *Sensor Review*, 32(4).

Liu, L. and Stamos, I. (2007). A systematic approach for 2D-image to 3D-range registration in urban environments. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. Ieee.

Mastin, A., Kepner, J., and Fisher III, J. (2009). Automatic registration of LIDAR and optical images of urban scenes. *Computer Vision and Pattern Recognition*, pages 2639–2646.

Napier, A., Corke, P., and Newman, P. (2013). Cross-Calibration of Push-Broom 2D LIDARs and Cameras In Natural Scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Nieto, J., Monteiro, S., and Viejo, D. (2010). 3D geological modelling using laser and hyperspectral data. *Geoscience and Remote Sensing Symposium*, pages 4568–4571.

Pandey, G., McBride, J., and Eustice, R. (2011). Ford campus vision and lidar data set. In *The International Journal of Robotics Research*, pages 1543–1552.

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

Pluim, J. P., Maintz, J. B., and Viergever, M. a. (2000). Image registration by maximization of combined mutual information and gradient information. *IEEE transactions on medical imaging*, 19(8):809–14.

Pluim, J. P. W., Maintz, J. B. A., and Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *Medical Imaging, IEEE*, 22(8):986–1004.

Rusu, R. B. (2010). Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. *KI - Künstliche Intelligenz*, 24(4):345–348.

Schneider, D. and Maas, H.-G. (2003). Geometric modelling and calibration of a high resolution panoramic camera. *Optical 3-D Measurement Techniques VI.*

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition*, pages 1–8.

Studholme, C., Hill, D. L., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern recognition*, 32(1):71–86.

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

Torabi, A. and Bilodeau, G. (2011). Local self-similarity as a dense stereo correspondence measure for themal-visible video registration. In *Computer Vision and Pattern Recognition*, pages 61–67.

Wachinger, C. and Navab, N. (2010). Structural image representation for image registration. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 23–30.